UNIVERSIDAD TECNOLÓGICA DE PEREIRA

DISEÑO Y CONSTRUCCIÓN DE UN TOOLBOX EN AMBIENTE SCILAB QUE APOYE LA ENSEÑANZA-APRENDIZAJE DE LA REGRESIÓN LINEAL Y ADEMÁS OFREZCA ALTERNATIVAS DE SOLUCIÓN AL PROBLEMA DE MULTICOLINEALIDAD.

JOSÉ DANIEL GALLO GALLÓN

UNIVERSIDAD TECNOLOGICA DE PEREIRA

FACULTAD DE CIENCIAS BASICAS

MAESTRIA EN ENSEÑANZA DE LAS MATEMATICAS

PEREIRA 2013

DISEÑO Y CONSTRUCCIÓN DE UN TOOLBOX EN AMBIENTE SCILAB QUE APOYE LA ENSEÑANZA-APRENDIZAJE DE LA REGRESIÓN LINEAL Y ADEMÁS OFREZCA ALTERNATIVAS DE SOLUCIÓN AL PROBLEMA DE MULTICOLINEALIDAD.

JOSÉ DANIEL GALLO GALLÓN

Trabajo final presentado como requisito para optar al título de
Magister en la Enseñanza de las Matemáticas con énfasis en Estadística

DIRECTOR
Magister Lorenzo Julio Martínez Hernández

UNIVERSIDAD TECNOLOGICA DE PEREIRA

FACULTAD DE CIENCIAS BASICAS

MAESTRIA EN ENSEÑANZA DE LAS MATEMATICAS

PEREIRA 2013

DEDICATORIA

Dedico este trabajo a los dos seres que más amo en la tierra, a mi Madre, que me dio la vida y luego me protegió de tal manera que me hace sentir un niño y a mi lindo y hermoso hijo, Juan Daniel que es mi dulce y alegre compañía en todos los aspectos.

AGRADECIMIENTOS

Deseo agradecer a

Dios y a mi Padre José Abel por todo lo que me brindan desde lo eterno

Mi Madre, María de los Ángeles Gallón

Mis hermanos que apoyaron

Mis compañeros del Departamento de Física y Matemáticas

Mi hijo, mi dulce y alegre compañía

RESUMEN

La regresión lineal en los cursos de pregrado es de mucha importancia para modelar fenómenos aleatorios. Con un buen modelo matemático se puede intentar describir el comportamiento de ciertos fenómenos que no son determinísticos y que nos pueden llevar a la toma de buenas decisiones en el campo profesional.

Los conceptos de regresión lineal aparecen en un contexto distinto al que se tiene para la función lineal, por tal motivo se hace necesario tener una capacidad amplia del comportamiento del fenómeno y de los datos. Para ello es indispensable recurrir a conceptos o expresiones matemáticas que adquieren significados distintos cuando se usan en el campo de la estadística. Se tiene un conjunto de puntos que expresan los valores que asumen dos características reales medibles y se quiere encontrar la relación que las vincula. Si bien el procedimiento que se utiliza para ello es de naturaleza determinística su interpretación a priori es de naturaleza aleatoria ya que distintos conjuntos de puntos, que representan a las mismas características, podrían originar distintas rectas. Cuestiones tales como parámetros, estimadores, variables aleatorias, independencia y distintas interpretaciones de la linealidad hacen su aparición en este tipo de problemas.

PALABRAS CLAVES

Regresión lineal, parámetro, estimadores, variables aleatorias, independencia estadística

ABSTRACT

Linear regression in undergraduate courses is very important to model random phenomena. With a good mathematical model can try to describe the behavior of certain phenomena that are not deterministic and that can lead to making good decisions in the professional field.

The linear regression concepts appear in a different context to that function are linear, so why it is necessary to have a large capacity phenomenon behavior and data. This is necessary to utilize math concepts or expressions that acquire different meanings when used in the field of statistics. It has a set of points that express the values they take on two real measurable characteristics and want to find the relationship that links them. Although the procedure used for this is by nature a priori deterministic interpretation is random in nature as different sets of points representing the same characteristics, may cause different lines. Issues such as parameters, estimators, random variables, independence and different interpretations of linearity make their appearance in this type of problems.

KEYWORDS

Linear regression parameter estimators, random variables, statistical independence

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

CAPÍTULO 1: INTRODUCCIÓN

La regresión y correlación lineal son temas de la estadística que ya forman parte del currículo de la mayoría de las carreras universitarias. Su inclusión posibilita el tratamiento y análisis de datos multifactoriales de la relación que puede existir entre las variables consideradas.

En la actualidad la estadística ocupa un lugar de gran importancia tanto en la investigación como en la práctica, y sin embargo esta situación es bastante reciente, basta señalar que el gran auge de la utilización del método estadístico, tanto para la planificación de experimentos como para el análisis de los datos obtenidos, podemos situarlos en los trabajos de quien sin lugar a dudas se considerado como el padre de la estadística moderna, Ronald A. Fisher (1890-1962).

Para el estudio de la estadística y en particular de la regresión lineal, es necesario el uso adecuado de un software estadístico o de algún programa computacional, que permita el manejo de gran cantidad de datos y la realización de un alto número de operaciones matemáticas en poco tiempo. Este tipo de recursos permite alcanzar con más facilidad el objetivo de la clase, aumentar el interés del tema por parte del estudiante, y la compresión de conceptos teóricos.

La regresión lineal es una técnica que se puede utilizar en cualquier área del conocimiento. Pero su enseñanza en el aula, sin un programa de cómputo es supremamente complejo, debido a la cantidad de operaciones algebraicas a realizar, a la cantidad de variables a tener en cuenta y a la gran cantidad de conceptos de suma importancia, en los que se debe profundizar para poder desarrollar un curso que valga la pena y no se quede en lo superficial. En la mayoría de las universidades, el docente sólo se centra en el método de mínimos cuadrados dado que no es posible profundizar en otros métodos.

En este trabajo se examinarán las asociaciones cuantitativas entre un número de variables por medio del método de los mínimos cuadrados y se tratara de buscar soluciones a ciertos problemas que éste método puede presentar en casos especiales.

Se trabajó permanentemente en los conceptos y la metodología básica para poder lograr, de una gran cantidad de datos el mayor provecho de las características asociada a estos, y poder predecir de la mejor manera para un futuro inmediato.

CAPÍTULO 2: OBJETIVOS

2.1 Objetivo General

Diseñar y construir un TOOLBOX EN AMBIENTE SCILAB que apoye la enseñanza-aprendizaje de la regresión lineal y que además ofrezca alternativas de solución al problema de multicolinealidad.

2.2 Objetivos Específicos

- Realizar un trabajo monográfico sobre la regresión lineal simple y múltiple y sus medidas remediales al problema de multicolinealidad.
- Diseñar material didáctico para la enseñanza aprendizaje de la estadística en el tema de regresión lineal a partir del uso del paquete de programación Scilab.
- Diseñar un tutorial en ambiente SCILAB que aporte al proceso de la enseñanza-aprendizaje de la regresión lineal con problemas prácticos.

CAPÍTULO 3: MARCO TEÓRICO

3.1 Historia de la regresión lineal

Legendre (1805) fue el primero en documentar el uso de la regresión lineal en una publicación del Método de los Mínimos Cuadrados que incluía una versión del teorema de Gauss-Márkov. Los primeros trabajos que tienen que ver con el estudio de la regresión lineal se remontan al siglo XIX, cuando Sir Francis Galton (1822-1917), estudio el impacto de la herencia en la estatura de las personas, y la expresión matemática de los fenómenos vinculados a ella. Él fue el primero en trabajar un conjunto de variables y asignar a la relación entre variables un número, para así obtener una medida referente a su grado de relación. Sostenía que las personas excepcionalmente altas solían tener hijos de estatura menor, mientras que las personas muy bajas solían tener hijos más altos; este hecho fue enunciado por Galton como la regresión a la media, aplicables a las tallas de una generación respecto de las siguientes. La justificación que se da hoy en día a esta situación es que los valores extremos de una distribución se deben en gran parte al azar [1].

3.2 Modelo lineal de dos variables (la recta)

Para tener una mejor comprensión en los conceptos de regresión y correlación lineal, revisaremos algunos términos de la fórmula de la ecuación de una línea recta.

$$y = \beta_0 + \beta_1 x \qquad (1)$$

El modelo de la ecuación lineal (1) que muestra la relación entre dos variables quedará representado gráficamente de forma general, así:

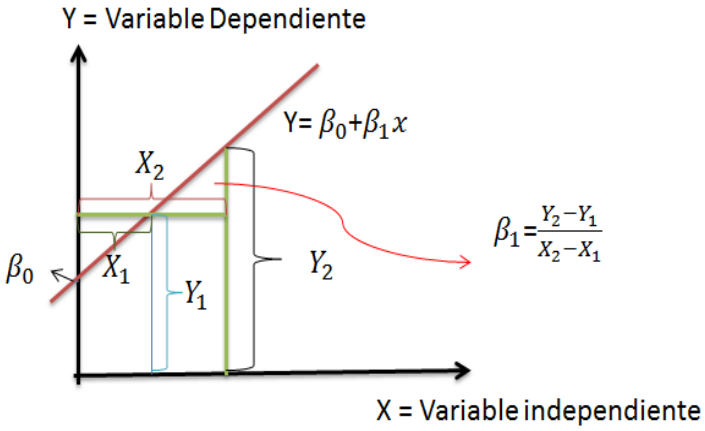


Figura Nº 1. La recta

En donde:

Y= Variable dependiente.

X = Variable independiente.

β_0= Punto de corte con el eje Y. llamada ordenada en el origen a sea que es el valor de Y cuando X es igual a cero

β_1= Pendiente de la recta. Es la variación en Y a causa de una variación en una unidad de la variable independiente.

3.3 Regresión lineal simple y formación del modelo

3.3.1 Introducción

El objetivo principal del análisis de regresión es determinar por medio de una ecuación lineal una relación cuantitativa entre la variable X (independiente) y la variable dependiente Y. Cuya finalidad principal es predecir el valor de la variable dependiente Y, a partir del o los valores de la o las variables independientes X.

El análisis de regresión lineal, es la técnica estadística más usada en la actualidad para realizar investigaciones y para modelar relaciones entre variables. En la regresión lineal simple hay una sola variable de regresión independiente X y una sola variable dependiente Y, y los datos se pueden representar mediante los pares $\{(X_i, Y_i); i =$

$1, 2, 3, \dots, n)\}$. Donde las variables X_i, Y_i, inicialmente se pueden llamar variable dependiente (Y), y Variable independiente (X).

Si la relación entre los pares de datos de X_i y Y_i es perfecta la ecuación o modelo, sería una ecuación lineal o recta mostrada en la ecuación (1).

Si observamos la ecuación (1), notamos que ésta posee dos parámetros β_0 y β_1, que vienen a ser los valores desconocidos de la ecuación que se deben encontrar en una investigación.

La regresión lineal es una técnica para investigar y modelar la relación entre variables. Se relaciona en gran medida con la estimación y/o predicción de la media (de la población) o valor promedio de la variable dependiente, con base en los valores conocidos o fijos de las variables explicativas X_i, (independientes).

Ni en la vida de los negocios, ni en los fenómenos físicos, los comportamientos son lineales exactos, ya que existen muchas variables que afectan el modelo. Por lo tanto el modelo de regresión lineal en estadística sufre una modificación y quedará de la siguiente forma:

$$y = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (2)$$

Como no es probable que todos los puntos estén exactamente sobre la línea de la recta, la ecuación lineal anterior (1), se debe modificar agregándole un término *de perturbación aleatoria, error, o término estocástico,* ε. como se ve en la ecuación (2).

La regresión determina si la variable (s) X e Y presentan una relación positiva, es decir, si las variables crecen a la vez, o si la relación es negativa porque dichas variables se desplazan en sentido opuesto. También determina la magnitud de la variación de Y para una variación dada de X.

3.3.2 Método de mínimos cuadrados

La regresión lineal en cualquier curso de estadística universitaria (pregrado), analiza el ***método de mínimos cuadrados***, que posee ciertos supuestos que se deben cumplir para poderlo usar adecuadamente, estos son:

- El modelo de regresión es lineal en los parámetros.
- Los valores de los regresores, las X, son fijos en muestreo repetido.
- Para X dadas, el valor medio de la perturbación ε_i es cero.
- Para X dadas, la varianza de $\mathcal{E}_i$ es constante ε homoscedástica.
- Para X dadas, no hay autocorrelación en las perturbaciones.
- Si las X son estocásticas, el término de perturbación y las X (estocásticas) son independientes o, al menos, no están correlacionadas.
- El número de observaciones debe ser mayor que el número de regresores.
- Debe haber suficiente variabilidad en los valores que toman los regresores.
- El modelo de regresión está correctamente especificado.
- No hay relación lineal exacta (es decir no hay Multicolinealidad).
- El término estocástico (de permutación) ε_i está normalmente distribuida [6].

El método de los mínimos cuadrados consiste en buscar los valores desconocidos de los parámetros $\hat{\beta}_0\ y\ \hat{\beta}_1$ de tal forma que la suma de los cuadrados de las diferencias entre observaciones (Y_i), y la línea recta calculada o estimada $\hat{Y}$, sea mínima. Así:

$$\sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \varepsilon = min. \qquad (3)$$

El termino $\left(Y_i - \hat{Y}\right) = \varepsilon$, es un error estadístico, que consiste en la diferencia que hay entre el valor observado de Y_i, y el de la línea recta $\hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 x)$, por lo tanto es una variable aleatoria que explica por qué el modelo no ajusta exactamente a los datos, y este error puede ser causado por los efectos de otras variables que no están analizados en el modelo.

Donde min, es el número más pequeño que se puede obtener si se suman estas desviaciones verticales elevadas al cuadrado $\sum_{i=1}^{n}(Y_i - \hat{Y})^2$, de aquí el nombre de método de los mínimos cuadrados para hallar la recta que mejor se ajusta, que da lugar a una recta que hace mínima la suma de los cuadrados de las distancias verticales desde cada punto observado Y_i, hasta la recta estimada $\hat{Y}$. Ver Figura Nº 2.

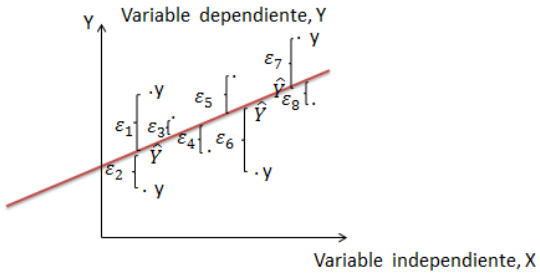


Figura Nº 2. Los residuales por mínimos cuadrados

Los residuales (ε), tienen un papel importante para investigar la adecuación del modelo de regresión ajustado y para detectar diferencias respecto a las hipótesis básicas [5].

Donde la variable X, que se le conocía como variable independiente, tomará el nombre de variable ***Predictora o regresora*** para no causar confusión con el concepto de independencia estadística y la variable Y, tomará el nombre de variable ***respuesta***.

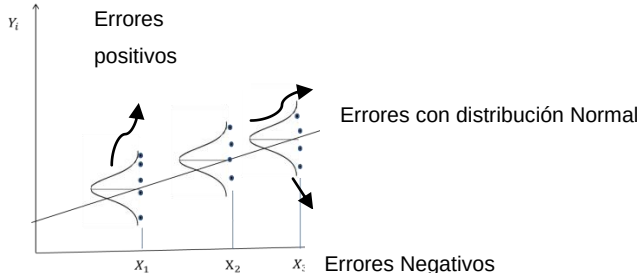


Figura Nº 3. Distribución de los ε_i

Para estimar estos parámetros por el método de mínimos cuadrados debemos partir de un modelo de regresión muestral esto es que está escrito en términos de los n pares de datos $(x_i, y_i), i = 1, 2, 3, \dots, n$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, 3, \dots, n \tag{4}$$

La ecuación (2) se considera como un modelo poblacional de regresión, mientras que la ecuación (4) es un modelo muestral de regresión, escritos en términos de n pares de datos.

Por lo anterior podemos suponer que hay una distribución de probabilidad de Y, para cada valor posible de X, con media condicional $Y_i = (Y|X_i) = \left(\mu_y \middle| X_i\right)$ (esperanza matemática) y varianza iguales a:

$$E(y|x) = \beta_0 + \beta_1 x \qquad (5)$$

La media condicional $E(y|x)$, o esperanza matemática nos dice que Y o variable respuesta es una variable aleatoria y que es una función lineal de X.

Es necesario recordar que el regresor X, se considera fijo, o lo que es igual, que el regresor X está controlado por el investigador o analista de datos.

$$Var(y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \qquad (6)$$

Donde $Var(y|x)$, es la varianza condicional (σ^2), y la varianza no depende de X.

Uno de los objetivos más importantes del análisis de regresión lineal simple es estimar los parámetros desconocidos $\beta_0\ y\ \beta_1$, del modelo de regresión. Estas estimaciones se denotarán como: $\hat{\beta}_0\ y\ \hat{\beta}_1$, que se leerán como beta sub-cero estimada para $\hat{\beta}_0$ y beta sub-uno estimada para $\hat{\beta}_1$, dado que para poder estimar estos valores tenemos que partir de una información muestral. Esta información consiste de un conjunto de pares ordenados de observaciones de X_i y Y_I, donde cada uno de estos pares pertenece a una unidad elemental particular de la muestra.

Para comprender el análisis y los criterios de la regresión por el método de los mínimos cuadrados debemos partir de:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \qquad (7)$$

Ahora, por derivadas parciales realizamos la primera derivada de (7) con respecto a β_0 e igualamos a cero para hallar una primera ecuación mínima (8), que será:

$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

Para hallar el mínimo igualamos a cero la ecuación última anterior, así:

$$-2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

Simplificamos

$$0 = \sum_{i=1}^{n} yi - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i$$

$$0 = \sum_{i=1}^{n} yi - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i \qquad (8)$$

Donde:

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} \qquad (9)$$

Como $\bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$, concluimos que:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \qquad (10)$$

Ahora para hallar β_1, derivamos la ecuación (8) con respecto a β_1, la igualamos a cero y hallamos una segunda ecuación, así:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_1} = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)[-(x_i)]$$

y

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0$$

Hallando una segunda ecuación.

$$-2\sum_{i=1}^{n} (x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0 \qquad (11)$$

Ahora se simplifican las ecuaciones (8) y (11) para obtener ***las ecuaciones normales*** de mínimos cuadrados que son:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \qquad (12)$$

A partir de estas ecuaciones normales de mínimos cuadrados se puede hallar las estimaciones de coeficientes de regresión $\hat{\beta}_0$y $\hat{\beta}_1$.

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad (13)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - \frac{(\sum_{i=1}^{n} Y_i)(\sum_{i=1}^{n} X_i)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} \qquad (14)$$

Donde el numerador de $\hat{\beta}_1$es la suma de los cuadrados X,Y.

$$SC_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), Suma\ de\ cuadrados\ del\ producto\ cruzado\ XY \qquad (15)$$

$$SC_{xy} = \sum_{i=1}^{n} Y_i X_i - \frac{(\sum_{i=1}^{n} Y_i)(\sum_{i=1}^{n} X_i)}{n} \quad (16)$$

Y el denominador de $\hat{\beta}_1$ es la suma de cuadrados de X.

$$SC_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \ Suma\ de\ cuadrados\ de\ X \quad (17)$$

$$SC_x = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} \quad (18)$$

Si tenemos en cuenta las ecuaciones (14) y (15) para estimar $\hat{\beta}_1$ la ecuación seria:

$$\hat{\beta}_1 = \frac{SC_{xy}}{SC_x} \quad (19)$$

Y también es necesario determinar la suma de cuadrados de Y, que es de mucha importancia para comprender el concepto de otras formulas posteriores; esta es:

$$SC_y = \sum_{i=1}^{n} (y_i - \bar{y})^2, \qquad Suma\ de\ cuadrados\ de\ Y \quad (20)$$

3.3.3 Propiedades de los estimadores por mínimos cuadrado

Los estimadores $\hat{\beta}_0\ y\ \hat{\beta}_1$ por mínimos cuadrados tienen algunas propiedades importantes [5].

Una primera propiedad, es que los estimadores $\hat{\beta}_{01}\ y\ \hat{\beta}_1$son combinaciones lineales de las observaciones Y_i. Según ecuaciones (13) y (14) se obtiene: [5].

$$\hat{\beta}_1 = \frac{SC_{xy}}{SC_x} = \sum_{i=1}^{n} C_i Y_i$$

Donde $C_i = \frac{(X_i - \bar{X})}{SC_x}$, $para\ i = 1, 2, 3 \ldots, n.$

Los estimadores de mínimos cuadrados son estimadores lineales insesgados óptimos.

Por lo anterior:

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} C_i Y_i\right) = \sum_{i=1}^{n} C_i E(Y_i)$$

$$= \sum_{i=1}^{n} C_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^{n} C_i + \beta_1 \sum_{i=1}^{n} C_i X_i$$

Ya que, se supuso que:

$E(\varepsilon_i) = 0$. Ahora se puede demostrar en forma directa que $\sum_{i=1}^{n} C_i = 0$ y que $\sum_{i=1}^{n} C_i X_i = 1$, y entonces

$$E(\hat{\beta}_1) = \beta_1$$

Esto es, si se supone el modelo es correcto (que $E(Y_i) = (\beta_0 + \beta_1 X_1)$, entonces $\hat{\beta}_1$es un estimador Insesgado de β_1. De igual manera se puede demostrar que $\hat{\beta}_0$ es un estimador Insesgado de β_0, es decir: [5].

$$E(\hat{\beta}_0) = \beta_0$$

La varianza de $\hat{\beta}_1$ se calcula como sigue:

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} C_i Y_i\right)$$

$$= \sum_{i=1}^{n} C_i^2 Var(Y_i) \qquad (21)$$

Ya que las observaciones de Y_i son no correlacionadas, por lo que la varianza de la suma es igual a la suma de las varianzas. La varianza de cada término en la suma de es $C_i^2 Var(Y_i)$ y hemos supuesto que $Var(Y_i) = \delta^2$; en consecuencia: [5].

$$Var(\hat{\beta}_1) = \delta^2 \sum_{i=1}^{n} C_i^2 = \frac{\delta^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{SC_x^2}$$

$$Var(\hat{\beta}_1) = \frac{\delta^2}{SC_x} = \frac{\sum_{i=1}^{n} \varepsilon^2}{n-k} * \frac{1}{\sum_{i=1}^{n} X_i^2} \quad (22)$$

La varianza de $\hat{\beta}_0$ es: $Var\left(\hat{\beta}_0\right) = Var\left(\bar{Y} - \hat{\beta}_1 \bar{X}\right)$

$$Var(\hat{\beta}_0) = \delta^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SC_x}\right) = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-k} * \frac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i} \quad (23)$$

Donde,

n = número de pares ordenados

k = número de parámetros a calcular

Para que los estimadores no tengan sesgo se debe cumplir con que la esperanza matemática de los $\hat{\beta}$ debe ser igual al parámetro β. Es decir: $E(\hat{\beta}) = \beta$. Por lo tato el sesgo = $E(\hat{\beta}) - \beta$.

Decir que los estimadores son óptimos o eficientes significa que su varianza es mínima. Los estimadores de mínimos cuadrados son estimadores óptimos o eficientes respecto a todos los estimadores insesgados. Esto se conoce como el teorema de *Gauss-Márkov* y representa la justificación más importante para utilizar los estimadores [8].

Existen otras propiedades útiles en el análisis de regresión por mínimos cuadrados que merecen ser nombradas:

La suma de los residuales en cualquier modelo de regresión que tenga una ordenada al origen $\hat{\beta}_0$ siempre es igual a cero.

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} \varepsilon_i = 0$$

La suma de los valores observados Y_i es igual a la suma de los valores ajustados $\hat{Y}_i$.

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

La línea de regresión por mínimos cuadrados siempre pasa por el centroide de los datos, que es el punto $(\bar{X}, \bar{Y})$.

La suma de los residuales, ponderados por el valor correspondiente de la variable regresora, siempre es igual a cero.

$$\sum_{i=1}^{n} X_i \varepsilon_i = 0$$

La suma de los residuales, ponderados por el valor ajustado correspondiente, siempre es igual a cero.

$$\sum_{i=1}^{n} \hat{Y}_i \varepsilon_i = 0$$

3.3.4 Estimación de la varianza (δ^2)

Para poder realizar pruebas de hipótesis y formar estimados de intervalos de regresión es necesario un estimador de δ^2.

Cuando no se conoce δ^2, se puede calcular de la suma de los cuadrados de los residuales, o suma de cuadrados del error.

$$SCE = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (23)$$

Se puede deducir una formula cómoda para SCE sustituyendo $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ en la ecuación (23) y simplificando.

$$SCE = \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} n\bar{Y}^2 - \hat{\beta}_1 SC_{xy} \qquad (24)$$

Pero

$$\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \equiv SCT \qquad (25)$$

Por lo anterior

$$SCE = SCT - \hat{\beta}_1 SC_{xy} \tag{26}$$

La suma de los cuadrados de los residuales tiene $n-2$ grados de libertad, porque los grados de libertad se asocian con los estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ que se usan para obtener $\hat{Y}_i$. Se ha demostrado que el valor estimado SCE es $E(SCE) = (n-2)\delta^2$, por lo que el estimador insesgado de δ^2 es

$$\hat{\delta}^2 = \frac{SCE}{n-2} = MSCE \tag{27}$$

Donde $MSCE$ se le llama cuadrado medio residual. Y a la raíz cuadrada de $\hat{\delta}^2$ se le llama, a veces error estándar de la regresión y tiene las mismas unidades que la variable de respuesta Y.

Ya que $\hat{\delta}^2$ depende de los cuadrados de la suma de los residuales, cualquier violación de las hipótesis sobre los errores del modelo, o cualquier especificación equivocada de la forma del modelo puede dañar gravemente la utilidad de $\hat{\delta}^2$ como estimador de δ^2. Como δ^2 se calcula con los residuales del modelo de regresión, se dice que es un estimado de δ^2 dependiente del modelo [5].

3.3.5 Prueba de hipótesis

Para realizar pruebas de hipótesis en regresión lineal es necesario hacer una hipótesis adicional que consiste en que los ε_i se distribuyen normalmente. Así, las hipótesis completas son: que los errores estén distribuidos normalmente y en forma independiente, con media cero y varianza δ^2.

Como en la mayoría de los casos se desconoce δ^2, el estadístico t, es un buen estadístico de prueba.

3.3.5.1 Prueba de hipótesis para probar si el coeficiente de correlación es diferente de cero.

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

$$t = \frac{r - \rho}{S_r} \tag{28}$$

Dónde:

t = Estadístico de Student.

S_r= Error estándar del coeficiente de correlación.

ρ= Coeficiente de correlación Poblacional Hipotético.

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} \tag{29}$$

n= Número de observaciones pareadas.

3.3.5.2 Pruebas de hipótesis para probar si los coeficientes $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de la regresión difieren de cero.

Para probar hipótesis acerca de la ordenada al origen, se puede utilizar

$$H_0: \beta_0 = 0$$

$$H_1: \beta_0 \neq 0$$

Y se podría usar el estadístico de prueba

$$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}} \tag{30}$$

donde

$$S_{\widehat{\beta}_0} = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-k} * \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_{i=1}^2}} \tag{31}$$

Para probar la pendiente

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Esta prueba se relaciona con la significancia de la regresión. El no rechazar $H_0: \beta_1 = 0$ implica que no hay relación lineal entre $X\ y\ Y$.

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{S_{\widehat{\beta}_1}} \tag{32}$$

$$S_{\widehat{\beta}_1} = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-k} * \frac{1}{\sum_{i=1}^{n} x_{i=1}^2}} \tag{33}$$

3.3.6 Intervalos de confianza y de predicción

En el mundo de la regresión lineal por método de mínimos cuadrados, las formulas más usadas en un curso normal, en resumen son las siguientes:

3.3.6.1 Intervalo de confianza: Es el intervalo que se usa para estimar el valor medio de $\hat{Y}$ para un valor especifico de X_p.

$$\hat{Y} \pm tS_{\hat{\mu}.x} \tag{34}$$

Donde

$$S_{\hat{\mu}.x} = S_{y.x}\sqrt{\frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \tag{35}$$

3.3.6.2 Intervalo de predicción: Es el intervalo que se usa para pronosticar un valor especifico de $\hat{Y}$, para un valor dado de X_i.

$$\hat{Y} \pm tS_{\hat{y}.x} \tag{36}$$

$S_{\hat{y}.x}$= Error estándar de la predicción.

$$S_{\hat{y}.x} = S_{y.x}\sqrt{1 + \frac{1}{n} + \frac{(X_p - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \tag{37}$$

3.3.7 Análisis de la varianza

También se puede usar el análisis de la varianza para probar el significado de la regresión. Este análisis se fundamenta en la partición de la variabilidad total de la variable Y de respuesta.

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right) \tag{38}$$

Si se elevan al cuadrado los dos lados de ecuación (38), y se suma para todas las n observaciones, se tiene

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}_i\right)(Y_i - \hat{Y}_i)$$

El tercer término del lado derecho de esta ecuación se puede escribir de la siguiente forma:

$$2\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}_i\right)(Y_i - \hat{Y}_i) = 2\sum_{i=1}^{n}\hat{Y}_i\left(Y_i - \hat{Y}_i\right) - 2\bar{Y}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$$

$$2\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}_i\right)(Y_i - \hat{Y}_i) = 2\sum_{i=1}^{n}\hat{Y}_i\varepsilon_i - 2\hat{Y}\sum_{i=1}^{n}\varepsilon_i = 0$$

Dado que la suma de los residuales siempre es igual a cero y la suma de los residuales ponderados por el valor ajustado $\hat{Y}_i$ correspondiente también es igual a cero, tenemos:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (39)$$

El lado izquierdo de la ecuación (39) es la suma corregida de cuadrados de la observaciones,SC_T, que miden la variabilidad total en las observaciones. Los dos componentes de SC_T miden, respectivamente, la cantidad de variabilidad en las observaciones Y_i explicada ($SC_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$) por la línea de regresión, y la variación residual que queda sin explicar ($SC_\varepsilon = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$) por la línea de regresión [5].

La ecuación (39) es fundamental para el análisis de la varianza para un modelo de regresión lineal y su forma simbólica será:

$$SC_T = SC_R + SC_\varepsilon \qquad (40)$$

Se puede aplicar el estadístico de prueba F normal del análisis de la varianza para probar $H_0: \beta_1 = 0$.

La fórmula del estadístico F es:

$$\mathrm{F} = \frac{SC_R/(\mathbf{K}-\mathbf{1})}{SC_\varepsilon/(\mathbf{n}-\mathbf{K})} = \frac{MS_R}{MS_\varepsilon} \qquad (41)$$

La tabla del análisis de la varianza es:

Tabla Nº 1: ANOVA. Análisis de la Varianza

FUENTE DE VARIACION	g.l	Suma de cuadrados	Estimación de las Var.	Coeficiente F.
Regresión	**k- 1**	$\boldsymbol{SC_R}$	$\boldsymbol{SC_R}/(\boldsymbol{k}-\mathbf{1})$	$\frac{MS_R}{MS_\varepsilon}$

Error residual	**n - k**	$\boldsymbol{SC_\varepsilon}$	$\boldsymbol{SC_\varepsilon}/_{(\boldsymbol{n-k})}$	
Total	**n - 1**	$\boldsymbol{SC_T}$		

Fuente [5].

$\boldsymbol{SC_R}$ **=** Suma de cuadrados de la regresión. $\sum_{i=1}^{n}(\hat{Y}-\bar{Y})^2$. Llamada también Variación explicada

$\boldsymbol{SC_\varepsilon}$ **=** Suma de cuadrados del error. $\sum_{i=1}^{n}(Y_i-\hat{Y})^2$. Llamada también Variación no explicada

$\boldsymbol{SC_T}$ **=** Suma de cuadrados total. $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$. Llamada variación total.

$$\boldsymbol{SC_T = SC_R + SC_\varepsilon}$$

K = número de parámetros linealmente independientes que se deben estimar ($\hat{\beta}_i$).

g.l = Grados de libertad.

3.3.8 Coeficiente de determinación

Este coeficiente mide el porcentaje de variabilidad en Y que puede ser explicado por la variable predictora X. Sus fórmulas más usadas son:

$$R^2 = \frac{SC_R}{SC_T} = 1 - \frac{SC_\varepsilon}{SC_T} \qquad (42)$$

3.4 Regresión lineal Múltiple.

Es un modelo de regresión donde intervienen más de una variable regresora. Así:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k + \varepsilon_i \qquad (43)$$

Se llama modelo de regresión lineal múltiple con k regresores. Los parámetros β_j, j = 0, 1, 2,…,k se llaman coeficientes de regresión. Este modelo describe un hiperplano en el espacio de k dimensiones de las variables regresoras X_j. El parámetro β_j representa el cambio esperado en la repuesta $\hat{Y}_i$ por un cambio unitario de X_j cuando todas las demás variables regresoras X_i $(i \neq j)$ se mantienen constantes. Por tal razón, a los parámetros β_j, j = 1, 2…k se les llama frecuentemente coeficientes de regresión parcial [5].

3.4.1 Estimación de los coeficientes de regresión por mínimos cuadrados

Se puede aplicar el método de mínimos cuadrados para estimar los coeficientes de regresión. Supongamos que $n > k$ observaciones, y sea Y_i la i-ésima respuesta observada, y X_{ij} la i-ésima observación o nivel de regresor X_j. Los datos aparecen en la tabla Nº 2

Tabla Nº 2: Datos para la regresión lineal múltiple

$\boldsymbol{Observación}$	$\boldsymbol{respuesta}$	$\boldsymbol{Regresores}$	
$\boldsymbol{i}$	$\boldsymbol{Y_i}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2 \dots X_k}$
1	Y_1	X_{11}	$X_{12} \dots X_{1k}$
2	Y_2	X_{21}	$X_{22} \dots X_{2k}$
⋮	⋮		⋮
n	Y_n	X_{n1}	$X_{n2} \dots X_{nk}$

El modelo de regresión se puede escribir de la siguiente forma:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} + \varepsilon_i$$

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ij} + \varepsilon_i\,, \quad i = 1, 2, 3, \dots n \tag{44}$$

En regresión lineal las variables regresoras $X_1, X_2, X_3, \dots X_k$, son fijas lo que indica que son variables no aleatorias y que se miden sin error.

La función de mínimos cuadrados es

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2 \qquad (45)$$

Como se debe minimizar la función S respecto a los parámetros $\beta_0, \beta_1, \dots, \beta_k$. sus estimadores por el método de mínimos cuadrados deben satisfacer

$$\frac{\partial S}{\partial \beta_0}\bigg| \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j) = 0 \qquad (46)$$

Y

$$\frac{\partial S}{\partial \beta_j}\bigg| \hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j) x_{ij} = 0\,, \quad j = 1, 2, \dots, k \qquad (47)$$

Se simplifican las ecuaciones anteriores (46 y 47) y encontramos las ecuaciones normales de mínimos cuadrados

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} X_{i2} + \cdots + \hat{\beta}_k \sum_{i1}^{n} X_{ik}$$

$$\sum_{i=1}^{n} X_{i1} Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} X_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{i1} X_{i2} + \cdots + \hat{\beta}_k \sum_{i1}^{n} X_{i1} X_{ik}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\sum_{i=1}^{n} X_{ik} Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} X_{ik} X_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} X_{i1} X_{i2} + \cdots + \hat{\beta}_k \sum_{i1}^{n} X_{ik}^2$$

Obsérvese que hay $p = k + 1$ ecuaciones normales, una para cada una de los coeficientes desconocidos de la regresión lineal y la solución de estas ecuaciones serán los estimadores por el método de los mininos cuadrados.

3.4.2 Notación matricial

Es más cómodo manejar modelos de regresión lineal múltiple cuando se presentan en notación matricial, para ello partamos de la fórmula para el modelo matricial.

$$Y = X\beta + \varepsilon \tag{48}$$

En donde

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Se desea determinar el vector $\hat{\beta}$ de estimadores de mínimos cuadrados que minimice

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' * \varepsilon = (Y - X\beta)'(Y - X\beta)$$

Que se puede expresar

$$S(\beta) = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta$$

$$S(\beta) = Y'Y - 2\beta'X'Y + \beta'X'X\beta$$

Dado que $\beta'X'Y$ es una matriz $1 * 1$ o sea un escalar, y su transpuesta $(\beta'X'Y)' = Y'X\beta$ es el mismo escala. Los estimadores de mínimos cuadrados deben satisfacer

$$\frac{\partial S}{\partial \beta}\Big| \widehat{\beta} = -2X'Y + 2X'X\hat{\beta} = 0$$

Si, se simplifica la ecuación anterior se llega a

$$X'X\hat{\beta} = X'Y \tag{49}$$

La ecuación matricial anterior (49) representa las ecuaciones normales de mínimos cuadrados vistas algébricamente.

Si multiplicamos la formula (49) por inversa de $X'X$, el estimador de β por mínimos cuadrados será:

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (50)$$

Siempre y cuando exista la matriz inversa $(X'X)^{-1}$. La matriz $(X'X)^{-1}$ siempre existe si los regresores son linealmente independientes, esto es, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

Para ampliar la parte conceptual de la regresión lineal de la forma matricial de las ecuaciones normales tenemos

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} & \cdots & \sum_{i=1}^{n} X_{ik} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} & \cdots & \sum_{i=1}^{n} X_{i1}X_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} X_{ik} & \sum_{i=1}^{n} X_{ik}X_{i1} & \sum_{i=1}^{n} X_{ik}X_{i2} & \cdots & \sum_{i=1}^{n} X_{ik}^2 \end{bmatrix} * \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1}Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{ik}Y_i \end{bmatrix}$$

Si se hace la multiplicación indicada se obtiene la forma escalar de las ecuaciones normales. Aquí podemos observar que $X'X$ es una matriz simétrica de $p * p$ y que $X'Y$ es un vector columna de $p * 1$. Tambien se puede ver la estructura especial de $X'X$. Los elementos diagonales de $X'X$ son las sumas de los cuadrados de los elementos de las columnas de X, y los elementos fuera de la diagonal son las sumas de los productos cruzados de los elementos de las columnas de X. Además se puede observar que los elementos de $X'Y$ son las sumas de los productos cruzados de las columnas de X por las observaciones Y_i.

El modelo ajustado de regresión que corresponde a los niveles de las variables regresoras $x' = [1,\ x_1, x_2, \dots, x_k]$ es

$$\hat{y} = x'\hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j\, x_j$$

El vector de valores ajustados $\hat{y}_i$ que corresponde a los valores observados y_i es

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY \qquad (51)$$

Donde $H = X(X'X)^{-1}X'$ es una matriz $n * n$ y se le suele llamar matriz de sombrero. Aplica el vector de valores observados en un vector de valores ajustados. La matriz de sombrero y sus propiedades desempeñan un papel central en el análisis de regresión.

La diferencia entre el valor observado Y_i y el valor ajustado $\hat{Y}_i$ correspondientes es el residual $\varepsilon_i = Y_i - \hat{Y}_i$. Los n residuales se pueden escribir cómodamente con notación matricial como sigue:

$$\varepsilon = y - \hat{y} \tag{52}$$

Hay otras formas de expresar el vector de los residuales ε, que pueden ser útiles en el estudio de regresión son:

$$\varepsilon = y - X\hat{\beta} = y - Hy = (I - H)y \tag{53}$$

3.4.3 Intervalos de confianza y de predicción de la regresión múltiple

Los intervalos de confianza y los de predicción de regresión lineal múltiple también juegan un papel sumamente importante cuando hablamos de varias variables regresoras. En esta parte se presentaran en forma breve los intervalos de confianza y de predicción.

3.4.3.1 Intervalos de confianza para la regresión lineal múltiple

Para poder construir los intervalos de confianza de los coeficientes de regresión β_j, se continúa suponiendo que los ε_i están distribuidos normal e independientemente, con promedio cero y varianza δ^2.

$$\hat{y}_0 - t\alpha_{/2,n-p}\sqrt{\hat{\delta}^2 x_0'(X' * X)^{-1}x_0} \leq E(y|x_0) \leq \hat{y}_0 + t\alpha_{/2,n-p}\sqrt{\hat{\delta}^2 x_0'(X' * X)^{-1}x_0} \tag{54}$$

La ecuación (54) es la generalización de intervalos para el caso de regresión múltiple.

3.4.3.2 Intervalos de predicción para la regresión

Los intervalos de predicción de regresión múltiple nos sirven como estimador puntual de una observación futura $(Y_{0)}$ y su fórmula será

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\delta}^2(1 + x_0'(X' * X)^{-1}x_0)} \leq Y_0 \leq \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\delta}^2(1 + x_0'(X' * X)^{-1}x_0)} \qquad (55)$$

La fórmula (55) es una generalización para observaciones futuras

3.4.4 Matriz de correlación

En regresión múltiple el primer paso que se debe realizar es el de identificar la variable dependiente y las variables independientes o predictoras que se van a tener en cuenta en el modelo.

Para seleccionar las variables predictoras en regresión múltiple en resumen se deben tener en cuenta las siguientes dos reglas:

Una variable Predictora debe tener una correlación fuerte con la variable dependiente.

Una variable Predictora no debe tener una correlación fuerte o demasiado alta con ninguna otra variable predictora [14].

Después de seleccionar las variables se pasa a tomar una muestra aleatoria y se registran todas las variables para cada elemento de la muestra y por último se identifica la relación entre las variables de predicción y la dependiente, y entre las propias variables de predicción. Esto se hace mediante el análisis de los datos en un programa de computadora que realice una matriz de correlación para las variables. El tamaño de la matriz depende del número de variables que se investigan.

3.4.5 Multicolinealidad

La multicolinealidad es un problema serio que se puede presentar en la estimación de regresión lineal múltiple y puede influir mucho en la utilidad del modelo de regresión. La multicolinealidad implica una dependencia aproximadamente lineal entre los regresores (variables independientes), que puede influir en forma dramática sobre la capacidad de estimar coeficientes de regresión. La presencia de multicolinealidad tiene una gran cantidad de efectos graves sobre los estimados de coeficientes de regresión por mínimos cuadrados.

Si no hay relación lineal entre los regresores (X_i) se dice que estos son ortogonales y se pueden hacer inferencias estadísticas con relativa facilidad como:

Identificación de los efectos relativos de las variables regresoras (X_i).

Predicción y/o estimación.

Selección de un conjunto adecuado de variables para el modelo

En la práctica en la mayor parte de las aplicaciones de regresión lineal múltiple los regresores no son ortogonales y no necesariamente la falta de ortogonalidad es un problema grave, pero en muchos casos los regresores tienen relación lineal perfecta que influyen en inferencias erradas. Cuando hay dependencia lineal casi perfecta, se dice que existe Multicolinealidad.

3.4.5.1 Multicolinealidad o Colinealidad

Estos dos términos, son muy conocidos e importantes en estudios estadísticos que tratan de la regresión lineal múltiple.

La multicolinealidad o Colinealidad se presenta, cuando existe una fuerte correlación entre las variables explicativas o regresoras en un modelo de regresión múltiple. La correlación debe ser fuerte entre estas variables ya que en la realidad, siempre existirá alguna correlación entre dos variables por ejemplo en el campo de la econometría, la no correlación entre estas variables es algo ideal y no real.

3.4.5.2 Clase de multicolinealidad

- **Multicolinealidad exacta**: se afirma que hay multicolinealidad exacta, cuando una o más variables regresoras X_i, son una combinación lineal de otra, es decir, que el coeficiente de correlación $r = 1 \, o \, r = -1$. Esto hace que el determinante de la matriz $X' * X$ tenga un determinante igual a cero, lo que indica que existe dependencia lineal entre las variables, y que la matriz es singular o invertible.

La multicolinealidad exacta se da cuando el rango es menor al número de columnas: $R_g(X) = r < K$

Cuando hay multicolinealidad exacta no se pueden estimar los parámetros del modelo de regresión múltiple; lo que se estima son combinaciones lineales de ellos que reciben el nombre de funciones estimables.

También se dice que existe multicolinealidad exacta, cuando la autocorrelación es mayor a 0,9. Esta autocorrelación es la correlación entre una variable, y ella misma atrasada uno o más periodos, (en series de tiempo). (G.Reitsch, 1997)

- **Multicolinealidad aproximada**: se dice que hay multicolinealidad aproximada, cuando una o más variables regresoras, no son exactamente una combinación lineal de la otra, pero su coeficiente de correlación entre estas variables es muy cercano a uno por lo tanto el determinante de la matriz $X' * X$ es muy cercano a cero.

Existen métodos para diagnosticar y para manejar la multicolinealidad. En este trabajo se profundizará sobre estos métodos y se buscaran posibles soluciones como por ejemplo trabajar la regresión de Ridge.

3.4.5.3 Fuentes de multicolinealidad

Si tomamos el modelo de regresión lineal múltiple de la forma: $y = X\beta + \varepsilon$; en donde y es un vector $n * 1$ de respuesta, X es una matriz $n * p$ de variables regresoras, β es un vector $p * 1$ de las constantes desconocidas y ε es un vector de $n * 1$ de errores aleatorios siendo $\varepsilon_i \sim NID(0, \delta^2)$. Será conveniente suponer que las variables regresoras y la

respuesta se han centrado y escalonado a longitud unitaria, en consecuencia, $X'X$ es una matriz de correlaciones de $p * p$, entre regresores y Xy es un vector de $p * 1$, de correlaciones entre los regresores y la respuesta.

Existen cuatro fuentes de multicolinealidad:

- **El método de recolección de datos que se empleo**

Cuando el analista muestrea sólo un subespacio de la región de los regresores definidos.

- **Restricciones en el modelo o en la población**

Cuando hay restricciones físicas en la población, habrá multicolinealidad independientemente del método de muestreo que se emplee. Con frecuencias se presentan restricciones en problemas donde intervienen procesos de producción o químicos, cuando los regresores son los componentes de un producto y ésos suman una constante. [5]

- **Especificación del modelo**

Esto sucede cuando dos o más regresores tienen dependencia casi lineal, y el tener esos regresores puede contribuir a la multicolinealidad.

- **Un modelo sobredefinido**

Cuando un modelo de regresión múltiple tiene más variables regresora que observaciones.

3.4.5.4 Efectos de la multicolinealidad

Los efectos de la multicolinealidad en los estimadores por el método de mínimos cuadrados suelen ser muy graves.

Una fuerte multicolinealidad da como resultado grandes varianzas y covarianzas de los estimadores de coeficientes de regresión por mínimos cuadrados. Esto implica que

distintas muestras tomadas con los mismos valores de X podrían ocasionar estimaciones muy diferentes de los parámetros del modelo.

3.4.5.5 Diagnostico de multicolinealidad

En teoría se han propuesto varias técnicas para detectar la multicolinealidad, a continuación se nombran algunas de ellas:

- ***Examen de la matriz de correlación***

Es una medida muy sencilla para detectar multicolinealidad, consiste en la inspección de los elementos r_{ij} no diagonales en $X' * X$. Si los regresores X_i y X_j son cuasi lineales dependientes $|r_{ij}|$ será próximo a la unidad.

- ***Factores de inflación de la varianza***

Los elementos diagonales de la matriz $C = (X' * X)^{-1}$ son muy útiles para detectar multicolinealidad. Como C_{jj}, el j-ésimo elemento de la diagonal de C se puede escribir de la forma $C_{jj} = (1 - R_j^2)^{-1}$, donde R_j^2 el coeficiente de determinación obtenido cuando se hace regresión de x_j respecto a los demás $p - 1$ regresores. Si x_j es casi ortogonal a los regresores restantes, R_j^2 es pequeño y C_{jj} es cercano a la unidad, mientras que si x_j es casi linealmente dependiente en algún subconjunto de los regresores restantes, R_j^2 es cuasi lineal y C_{jj} es grande. Como la varianza de los j-ésimos coeficientes de regresión es $C_{jj}\delta^2$ se puede considerar que C_{jj} es el factor en el que aumenta la varianza de $\hat{\beta}_j$ debido a dependencias casi lineales entre los regresores. Tenemos [5].

$$VIF_j = C_{jj} = (1 - R_j^2)^{-1} \qquad (56)$$

El factor VIF (*de variance inflation factor*) para cada término del modelo mide el efecto combinado que tienen las dependencias entre los regresores sobre la varianza de ese término. Si hay uno o más VIF grandes hay multicolinealidad. La experiencia indica que si cualquiera de los VIF es mayor que 5 o 10, es indicio de que los coeficientes asociados de regresión están mal estimados debido a la multicolinealidad [5].

- ***Análisis del eigensistema de $X' * X$***

Las raíces características, o valores propios de $X' * X$, se pueden usar para medir el grado de multicolinealidad en los datos. Si hay una o más dependencias casi lineales en los datos, una o más de las raíces características será pequeña. Uno o más valores propios (λ_j) pequeños implican que haya dependencia casi lineal entre las columnas de X. Algunos analistas prefieren examinar el número de condición de $X' * X$, que se define como

$$k = \frac{\lambda_{máx}}{\lambda_{mín}} \qquad (57)$$

Donde: $\lambda_{máx}\ y\ \lambda_{mín}$, son los valores propios máximo y mínimo respectivamente de $X' * X$

Que no es más que una medida de dispersión en el espectro de valores propios de $X' * X$. En general, si el número de condición es menor que 100, no hay problema grave de multicolinealidad. Los números de condición de 100 a 1000 implican multicolinealidad moderada a fuerte, y si k es mayor a 1000 es indicio de una fuerte multicolinealidad [5].

3.4.5.6 Métodos para manejar la multicolinealidad

Para manejar problemas de multicolinealidad se han propuesto varias técnicas; entre las más generales esta la recolección de más datos, la reespecificación del modelo y el uso de métodos de estimación distintos de los mínimos cuadrados, diseñados en forma específica para combatir los problemas de multicolinealidad.

En este trabajo se muestra y estudia el método de **Ridge** para solucionar problemas de multicolinealidad.

3.4.6 Regresión de Ridge

Si se aplica el método de regresión de mínimos cuadrados a datos no ortogonales, los estimadores encontrados suelen ser muy malos, por ejemplo los estimadores de los coeficientes pueden estar muy inflados y la longitud del vector de los estimados de los parámetros por mínimos cuadrados es excesiva, en promedio. Eso implica que el valor

absoluto de los estimados por mínimos cuadrados sea demasiado grande y que esos estimados son muy inestables, indicando con esto que sus magnitudes y signos pueden cambiar mucho con una muestra distinta.

Este método se basa en la premisa de que el valor reducido del determinante de la matriz $X' * X$ puede causar problemas a la hora de aplicar el método de los mínimos cuadrados ordinarios. Para solucionar este problema existe un método llamado el estimador Ridge que consiste en sumar una determinada cantidad a los elementos de la diagonal principal de $X' * X$. En forma específica el estimador de Ridge ($\hat{B}_{Ridge}$) se define como la solución de

$$(X' * X + kI)\hat{B}_{Ridge} = X' * y$$

Que es

$$\hat{B}_{Ridge} = (X' * X + kI)^{-1}X' * y \qquad (58)$$

K, en la ecuación (58) es una constante que selecciona el analista. Y debe ser $k \geq 0$.

Existen métodos para la elección del parámetro k como la elección por inspección de la traza ridge que es un procedimiento subjetivo que requiere criterio por parte del analista. Varios autores han propuesto procedimientos para elegir k, que son más analíticos, así, por ejemplo:

Hoerl, Kennard y Baldwin [1975] sugieren que una elección adecuada de k es

$$k = \frac{p\hat{\delta}^2}{\hat{\beta}'\hat{\beta}} \qquad (59)$$

Donde la varianza ($\hat{\delta}^2$) y los betas estimados ($\hat{\beta}$) son determinados por el método de los mínimos cuadrados. Basados en la metodología de superficie de respuesta en diseño experimental se determina el valor de p llamado diseño compuesto central. Este valor se usa para ajustar una superficie de respuesta de segundo orden. Se define como:

$$p = 2\gamma + C_2^{\gamma}$$

Donde γ es el número de variables regresoras, y C_2^{γ} es una combinatoria.

Demostraron, con simulaciones, que el estimador ridge resultante tuvo mejora importante en el MSE respecto al de mínimos cuadrados.

McDonald y Galarneau [1975] sugieren escoger k de tal modo que:

$$\hat{\beta}'_{Ridge}\hat{\beta}_{Ridge} = \hat{\beta}'\hat{\beta} - \hat{\delta}^2 \sum_{j=1}^{p} \left(\frac{1}{\lambda_j}\right) \qquad (60)$$

Para los casos en que el lado derecho de la ecuación (60) es negativo, investigaron igualar $k = 0$ (minimos cuadrados) o $k = \infty$ ($\hat{\beta}_R = 0$). Ningún método, en todos los casos fue mejor que el de mínimos cuadrados [5].

Los métodos que se han usado para elegir k se enfocan en la mejora de los estimados de los coeficientes de regresión. Si el modelo se va a usar para predicción, será más adecuado tener en cuenta criterios orientados a las predicciones, para elegir k. Mallows [1973] modifico el estadístico C_p para formar un C_k que se puede usar para determinar k. Propuso graficar C_k en función de V_k, siendo

$$C_k = \frac{SS_{Res}(k)}{\hat{\delta}^2} - n + 2 + 2Tr(XL) \qquad (61)$$

$$V_k = 1 + Tr(X'XLL') \qquad (62)$$

$$L = (X'X + kI)^{-1}X' \qquad (63)$$

Donde $SS_{Res}(k)$ es la suma de cuadrados residuales en función de k. La sugerencia es escoger la k que minimice C_k. Obsérvese que

$$XL + X(X'X + kI)^{-1}X' \equiv H_k \qquad (64)$$

Y que H_k equivale a la matriz sombrero en mínimos cuadrados ordinarios.

Procedimiento $PRESS_{Ridge}$ en donde interviene $PRESS_{Cresta} = \sum_{i=1}^{n} \left(\frac{\varepsilon_{i,k}}{1-h_{ii,k}} \right)^2$ en donde $\varepsilon_{i,k}$ es el i-ésimo residual para un valor determinado de k, y $h_{ii,k}$ es el i-ésimo elemento diagonal de H_k. El valor de k se escoge de tal modo que minimice $PRESS_{Ridge}$. Nótese que este procedimiento sólo es un aproximación al valor verdadero de $PRESS_{Ridge}$ que se podría obtener en realidad eliminando una por una las observaciones (recalculando cada vez los estimados ridge), porque cuando se centran y escalan los datos, la eliminación de un punto de dato cambia las constantes de centrado y escalado, y en consecuencia las observaciones, sin embargo, si no hay grandes diagonales en la matriz de sombrero (puntos influyentes) y el tamaño de muestra no es pequeño, $PRESS_{Ridge}$ es una buena aproximación a $PRESS$ $\boldsymbol{exacto}$:

$$PRESS(k) = \sum_{i=1}^{n} \varepsilon_{(i),k}^2$$

En donde $\varepsilon_{(i),k}^2$ es el residual obtenido realmente conservando la i-ésima observación para un k particular, centrando y escalando los datos, ajustando el modelo de ridge y calculando $\hat{y}_{(i),k}$; por consiguiente $\varepsilon_{(i),k} = y_i - \hat{y}_{(i),k}$. Existe un programa de computo (PROCIML, SAS institute [1987] para graficar $PRESS(k)$ en función de k.

Wahba, Golub y Health [1979] sugirieron el estadístico generalizado de validación cruzada

$$GCV = \frac{\sum_{i=1}^{n} \varepsilon_{(i),k}^2}{\{n - [1 + Tr(H_k)]\}^2}$$

Al escoger k se selecciona de tal modo que se minimiza el estadístico GCV. Hay una relación obvia con los procedimientos análogos a PRESS, descritos arriba [5].

Hay muchas otras posibilidades de escoger k. Por ejemplo Marquardt [1970] propuso usar un valor de k tal que VIF quede entre 1 y 10, de preferencia más cercano a uno. Dempster, Shatzoff y Wermuth [1971], Goldstein y Smith[1974], Lawless y Wang [1976], Lindley y Smith [1972] y Obenchain [1975] propusieron otros métodos de selección de k.

3.5 Scilab

Scilab es un software matemático, creado en el año 1.990 por investigadores de INRIA (instituto nacional de investigación en informática y automática de Francia), que es una de las organizaciones líderes en el mundo en la transferencia e innovación tecnológica y por École nationale des ponts et chaussées (ENPC).

La Ecole des Ponts ParisTech es una gran escuela francesa que forma ingenieros de alto potencial de futuros líderes y de alto nivel de perfil científico y técnico, llamado a abordar los principales retos de la sociedad de hoy y de mañana. Ouverte aux étudiants et chercheurs de nombreuses nationalités, c'est une institution à taille humaine à l'ambiance multiculturelle et pluridisciplinaire. Abierto a estudiantes e investigadores de muchas nacionalidades, es una institución a escala humana con el medio ambiente multicultural y multidisciplinario [16].

En el año 2.003 se creó el consorcio Scilab (Scilab consortium) para ampliar y promover el software como una referencia para todo el mundo académico e industria y en julio del 2.008, con el fin de mejorar la trasferencia de tecnología, el Scilab consortium se unió a la Fundación Digiteo.

“Scilab viene con numerosas herramientas: gráficos 2-D y 3-D, animación, álgebra lineal, matrices dispersas, polinomios y funciones racionales, Simulación: programas de resolución de sistemas de ecuaciones diferenciales (explícitas e implícitas), Xcos: simulador por diagramas en bloque de sistemas dinámicos híbridos, Control clásico, robusto, optimización LMI, Optimización diferenciable y no diferenciable, Tratamiento de señales, Grafos y redes, Scilab paralelo empleando PVM, Estadísticas, Creación de GUIs, Interfaz con el cálculo simbólico (Maple, MuPAD), Interfaz con TCL/TK.

Además se pueden agregar numerosas herramientas o toolboxes hechas por los usuarios como Grocer una herramienta para Econometría u Open FEM (Una caja de Herramientas para Elementos Finitos), hecha por INRIA.

En el pasado Scilab podía ser utilizado en el análisis de sistemas, pero no podía interactuar con el exterior. Hoy en día se pueden construir interfaces para que desde Scilab se pueda manejar un dispositivo, se conecte a la red a través de Tcp (Protocolo de

Control de Transmisión) o Udp (User Datagram Protocol), etc. Esto brinda la posibilidad de conectar una placa de adquisición de datos a Scilab y de esta forma el control de una planta on-line" [2].

Scilab es un lenguaje de programación de alto nivel, para cálculo científico, interactivo de libre uso y disponible en múltiples sistemas operativos (Mac OS X, GNU/Linux, Windows) desarrollado por INRIA (Institut National de Recherche en Informatique et Automatique) y la ENPC (École Nationale des Ponts et Chaussées) desde 1990. Scilab es ahora desarrollado por Scilab Consortium dentro de la fundación Digiteo [2].

Scilab fue creado para hacer cálculos numéricos aunque también ofrece la posibilidad de hacer algunos cálculos simbólicos como derivadas de funciones polinomiales y racionales. Posee cientos de funciones matemáticas y la posibilidad de integrar programas en los lenguajes más usados (Fortran, Java, C y C++). La integración puede ser de dos formas; por ejemplo, un programa en Fortran que utilice Scilab o viceversa. Scilab fue hecho para ser un sistema abierto donde el usuario puede definir nuevos tipos de datos y operaciones entre los mismos [2].

3.6 Enseñanza

Dentro de muchas de las preocupaciones de quienes ejercen la enseñanza de la estadística, está el desarrollo de técnicas, y estrategias de enseñanza-aprendizaje que tengan en cuenta el uso de nuevas tecnologías que permitan alcanzar los objetivos propuestos.

Este trabajo tiene como FINALIDAD PROFUNDIZAR la comprensión de los conceptos más relevantes de la regresión y correlación lineal basada en el uso de un programa de computación, como lo es el Scilab. Este software permite realizar muchos cálculos matemáticos que a mano serían muy tediosos y complejos.

Con esta propuesta se pretende estructurar un curso de regresión lineal donde el software libre Scilab intervenga como facilitador del proceso enseñanza-aprendizaje. Es necesario modificar las estructuras curriculares, organizadas hoy en día a partir de

contenidos temáticos y centrados en el trabajo de lápiz y papel, hacia la búsqueda del desarrollo intelectual que incorpore las tecnologías informáticas con miras a fortalecer las actividades cognitivas. Ya que la sociedad contemporánea depende, para su desarrollo, de sus capacidades para producir, aplicar y transmitir el conocimiento científico y tecnológico. Estar en posesión de tales capacidades conlleva la producción de recursos humanos con una amplia y variada formación científica y humanística.

En busca de la profundización de muchos conceptos de suma importancia en la regresión lineal como por ejemplo su validez, comprobación de los supuestos de normalidad y la toma de medidas remédiales

Los educadores de hoy tenemos que proporcionar a las futuras generaciones herramientas que le permitan enfrentarse a la resolución de problemas, no sólo en el ámbito escolar sino en sus posibles lugares de trabajo, en donde la creatividad y la innovación será la moneda de cambio. Tenemos el reto de proporcionarles instrumentos de aprendizaje, es decir, estructuras cognitivas con alto grado de adaptación a lo nuevo [3].

El desarrollo y análisis de datos requiere de apoyo computacional. En la actualidad el manejo de gran cantidad de información, la rapidez que se necesita para dar respuesta a una situación hace indispensable y necesario el uso de un paquete computacional.

CAPÍTULO 4: ANTECEDENTES

De acuerdo con la revisión realizada para el presente trabajo, se pueden identificar algunos estudios realizados en el contexto latinoamericano que pueden aportar elementos para esta tesis:

Huertas [18], elaboró una investigación sobre las nuevas tecnologías en la didáctica estadística. Dicho trabajo pretende elaborar un WebQuest siguiendo seis etapas: introducción, tarea, proceso, recursos, evaluación y conclusión, dirigida a los estudiantes que se encuentran cursando la asignatura de de "Estadística e Introducción a la Econometría" de la Licenciatura en Administración y Dirección de Empresa. A partir de allí se trabajan los contenidos referentes al análisis estadístico de variables aleatorias bidimensionales, se le plantea al alumno la búsqueda de datos oficiales en las páginas web del Instituto Nacional de Estadística (INE) y del Instituto Andaluz de Estadística (IEA). Con los datos que obtiene de esta búsqueda se le pide al estudiante que realice un estudio descriptivo de las variables consideradas usando esos datos. Entre otras actividades, se les pide calcular las tablas de frecuencia, las medidas de centralización y dispersión de las variables, y el estudio de regresión y correlación.

Moreno Echavarría [19] presentó una forma de cómo enseñar el concepto de regresión lineal simple a estudiantes cuyos conocimientos estadísticos previos sólo son descriptivos, se basa en la teoría constructivista a partir de inducir al estudiante es a través de preguntas y una serie de pasos para la construcción del aprendizaje del modelo de regresión lineal, a partir de una situación problemática real planteada que permite que el estudiante en la búsqueda de la construcción del aprendizaje adquiera el concepto de regresión lineal simple, identifique los elementos matemáticos y estadísticos que lo componen y pueda generar predicciones a partir de la construcción de éste, así mismo pueda poner en práctica, ante otras situaciones planteadas, la aplicación del concepto adquirido. Unas de las conclusiones obtenidas por Moreno Echavarría es poder reconocer que la construcción de los modelos de regresión lineal se puede generar a partir de

preguntas, pero es necesario partir de los conocimientos previos de los estudiantes en estadística descriptiva.

También es importante tener en cuenta en la enseñanza de la estadística la importancia de la utilización de recursos didácticos y herramientas computacionales para la tabulación y análisis estadístico de datos, que proporcionen al estudiante la comprensión de los conceptos estadísticos. A través de situaciones problemáticas basadas en un contexto real los estudiantes pueden aprender con mayor facilidad el concepto de regresión lineal simple, ya que esto permite que el estudiante sea introducido a una experiencia de aprendiz en su propia formación [19].

Cirilo y Molina [20] dicen que la utilización de las redes virtuales como soporte de variadas experiencias dentro del mundo educativo generan nuevos entornos virtuales de aprendizaje en los cuales las TIC participan ampliamente. En dichos entornos se favorece la interactividad, se estimulan estrategias de comunicación y colaboración asincrónica y sincrónica, se facilita la comunicación a distancia, se propician las tareas referidas a hacer más accesible, editable y publicable la información compartida. Esta investigación es tomada como antecedente en tanto presenta la manera como se puede implementar entornos virtuales de aprendizaje de la matemática.

Molinet, Martínez y Casas [21] hablan de la enseñanza de la estadística involucrando la tecnología y lenguajes de programación. Existen investigaciones en las que se pretende utilizar un lenguaje de programación para el aprendizaje, de los métodos probabilísticas. De acuerdo con los autores, consideran que los métodos probabilísticos son utilizados en la generación de números aleatorios y que al usar la programación como herramienta, se puede visualizar mejor las aplicaciones de las distribuciones. Según los investigadores, la programación de estas distribuciones, empleando cualquier lenguaje de programación, permite al estudiante simular el comportamiento de sistemas y facilita la comprensión de procesos y fenómenos que serán cotidianos en su labor profesional y a su vez el estudio de las distintas condiciones límites a las que podrían verse sujetos dichos procesos, que abarcan desde lo industrial hasta lo social y biológico. La automatización de los cálculos permite que el estudiante resuelva en un tiempo razonable ejemplos de procesos reales, sin importar cuán extenso y complejos sean los cálculos, además de dotarlo de gráficas que faciliten la toma de decisiones. Dentro de los

resultados de este trabajo [21], se puede afirmar que el empleo de la programación representa una herramienta en los métodos probabilísticos para la generación de números aleatorios y sus aplicaciones en la simulación. Para la elaboración de esta tesis, el trabajo de Molinet, Martínez y Casas puede ser un antecedente debido a que incorpora en la enseñanza de la estadística, los lenguajes de programación y reconoce que se convierte en una poderosa herramienta para la visualización de procesos estocásticos [21].

Contreras, Molina, Arteaga [22], emplean el lenguaje R como una introducción a la estadística para profesores. R es un lenguaje y entorno de programación, creado en 1993 por Ross Ihaka y Robert Gentleman del Departamento de Estadística de la Universidad de Auckland, cuya característica principal es que forma un entorno de análisis estadístico para la manipulación de datos, su cálculo y la creación de gráficos En su aspecto R puede considerarse como otra implementación del lenguaje de programación S, con la particularidad de que es un software GNU, General Public Licenc (conjunto de programas desarrollados por la Free Software Foundation), es decir, de uso libre. Con este programa se puede trabajar regresión lineal y se convierte en antecedente para este estudio en cuanto muestra la posibilidad de incorporar tecnología en la enseñanza de la estadística [22].

Torres y Gilbert [23] elaboraron un estudio que da cuenta de un proyecto, en el que se pretende el uso de programas de aplicación en matemáticas y con miras a que sea utilizado para la comprensión de la Probabilidad y Estadística como una prioridad para las Escuelas desde los niveles Medio Superior y Superior para los ciclos escolares donde se imparten las enseñanzas de las Matemáticas. Se plantea como hipótesis que en la implementación de herramientas, el estudiante pondrá en práctica los métodos y conocimientos adquiridos en el aula, visualizando los resultados e interpretando los mismos de una manera amigable. Pondrán en práctica sus conocimientos adquiridos en cursos de computación y el uso de paquetes como statgraphics, Excel, statistica, Minitab, MAPLE entre otros.

CAPÍTULO 5: METODOLOGÍA

Para dar cumplimiento a los objetivos específicos planteados se realizó lo siguiente:

La revisión bibliográfica fue el primer paso, en ésta, se consultaron los conceptos teóricos más importantes que tienen que ver con un curso de regresión lineal en el aula de clase de una carrera de pregrado fundamentalmente. Se revisaron documentos que trataron sobre enseñanza-aprendizaje. También se recolectó información sobre el método de mínimos cuadrados su aplicación, sus ventajas y desventajas. Dentro de esta revisión bibliográfica sobre las dificultades de la enseñanza aprendizaje de la regresión lineal en los cursos de estadística II, se encontraron dificultades como la cantidad de operaciones repetitivas que el estudiante requería para hallar una ecuación de regresión. Luego se hizo un análisis de los conceptos fundamentales sobre regresión lineal que a la vista de varios programas de estadística II estudiados, eran relevantes para un futuro profesional que requiere tal conocimiento.

Se recopilo información sobre didáctica de las matemáticas, en especial sobre la Estadística, con el fin de tener los argumentos adecuados y pertinentes para la realización de este trabajo

En una primera instancia se trabajó buscando una estrategia que se dedicara a la solución de ciertos problemas que el autor por su experiencia laboral había observado, para ello se pensó en elaborar un toolbox como propuesta de enseñanza aprendizaje, mediante el uso de un paquete de programación. Es así, como aparece la ida de trabajar con SCILAB, ya que a raíz de la investigación se supo que era un paquete computacional similar al Matlab y que era un software libre.

Se estudiaron los principales comandos y algoritmos de SCILAB, sus aplicaciones y en general su uso. De tal manera que permitieran su aplicación y comprensión. Aquí nos dimos cuenta que para lograr el objetivo general del presente trabajo era necesario tener buenos conceptos de programación estructurada y de diseño de interfaces graficas. Dado que se quería un programa interactivo y de fácil digitación.

Como segundo paso se consultó acerca de la forma en que se crea una Toolbox (Caja de Herramientas) en Scilab. La mayoría de detalles técnicos se obtuvieron de wiki.scilab.org [24].

Como tercer y cuarto pasó se procedió respectivamente a realizar algunos programas en Scilab y a escribir un tutorial de regresión lineal. Las actividades de programar y redactar se realizaron de manera alternada. Primero se realizaron algunos programas para la solución de problemas de regresión lineal simple mediante instrucciones de código básicas. Así mismo se redactaron estos programas de manera detallada en el tutorial. Se describieron cada una de las funciones, operadores y detalles del código necesarios para la solución de cada problema. Luego se procedió a programar la Toolbox, primero las funciones relacionadas con regresión lineal simple y finalmente las de regresión múltiple. En el tutorial se hizo una explicación muy extensa y detallada de cada una de las funciones de la Toolbox. La explicación de cada función incluye una descripción, una secuencia de llamado y un ejemplo. La Toolbox incluye una interfaz gráfica en la que se puede, de manera interactiva, modificar datos, realizar gráficos, y observar datos estadísticos relacionados con la regresión lineal simple [25]. También en esta fase se diseño y se laboro un material didáctico para la enseñanza de la estadística en el tema de regresión lineal a partir del uso del paquete de programación Scilab. Ver Anexo Nº 1.

Para lograr superar los problemas de la enseñanza aprendizaje de la regresión lineal se estudiaron problemas prácticos de regresión lineal simple y múltiple con características y propiedades diferentes, con la finalidad de llevarlos al programa SCILAB y crear un material didáctico adecuado y poder diseñar un tutorial en ambiente SCILAB que aporte al proceso de la enseñanza-aprendizaje de la regresión lineal y sus medidas remediales de multicolinealidad.

En resumen la realización de este trabajo se divide en tres fases:

Fase Inicial:

Se realizó un trabajo bibliográfico sobre la regresión lineal simple y múltiple, el método de mínimos cuadrados, el problema de multicolinealidad y sus posibles soluciones para la regresión lineal múltiple. Se hizo una revisión exhaustiva en torno a los conceptos

de regresión lineal, su evolución histórica, las dificultades en su enseñanza y en su aplicación.

Fase Intermedia:

En esta fase se dedica la mayor parte del tiempo en el diseño del toolbox en ambiente SCILAB, y para ello se realizaron las siguientes actividades:

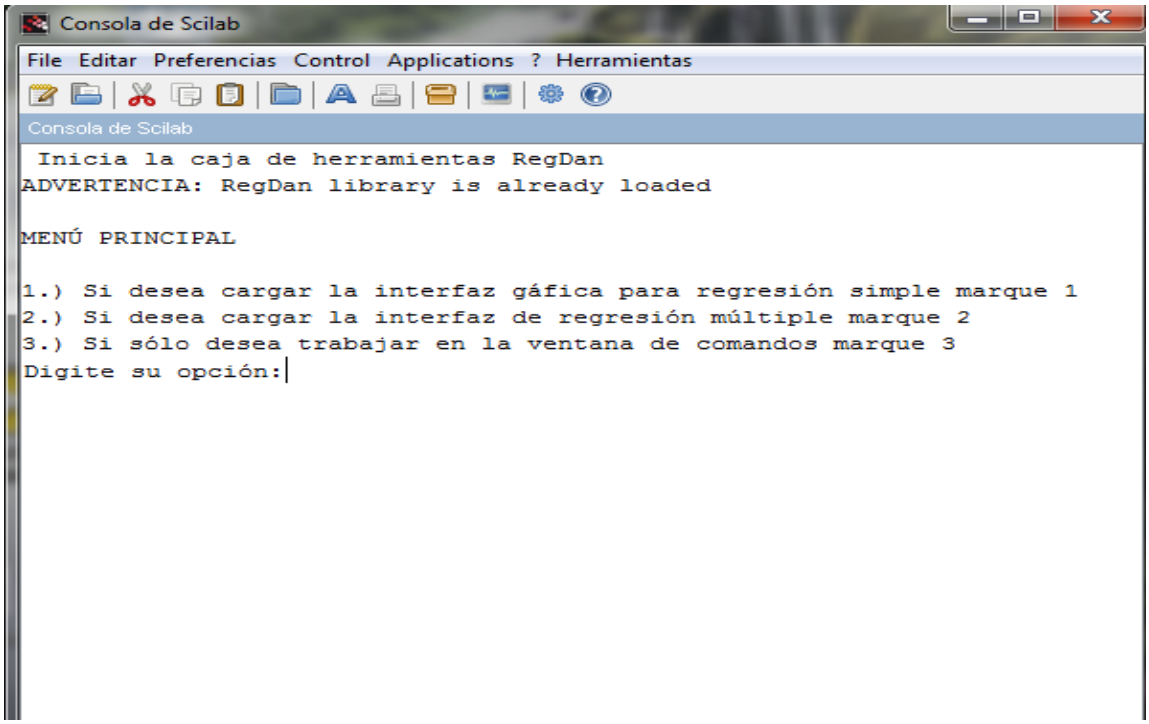
Para el presente trabajo de grado, se realizaron una serie de actividades bibliográficas sobre el programa Scilab. Ya que se trata de un software libre y con fines netamente académicos, se hizo muy fácil la descarga del programa y la obtención de la documentación existente. En la página web de la aplicación (www.scilab.org) se pudo descargar gratuitamente el programa para múltiples sistemas operativos, además se cuenta con una amplia gama de documentación al respecto. También se diseñó el Toolbox RegDan.

Fase Final:

Esta fase se puede llamar como de evaluación del proyecto, aquí el autor hace uso práctico del Toolbox RegDan, en dos grupos de estadística II (28 estudiantes), de la Universidad Autónoma de Manizales y lo pone a consideración de 7 docentes de departamento de física y matemáticas. Quienes al final resuelven una prueba de Likert que es una escala ordinal que midió la actitud de favorabilidad o desfavorabilidad del Toolbox RegDan en la enseñanza aprendizaje de la regresión simple, múltiple y la regresión de Ridge como medida remedial a la multicolinealidad.

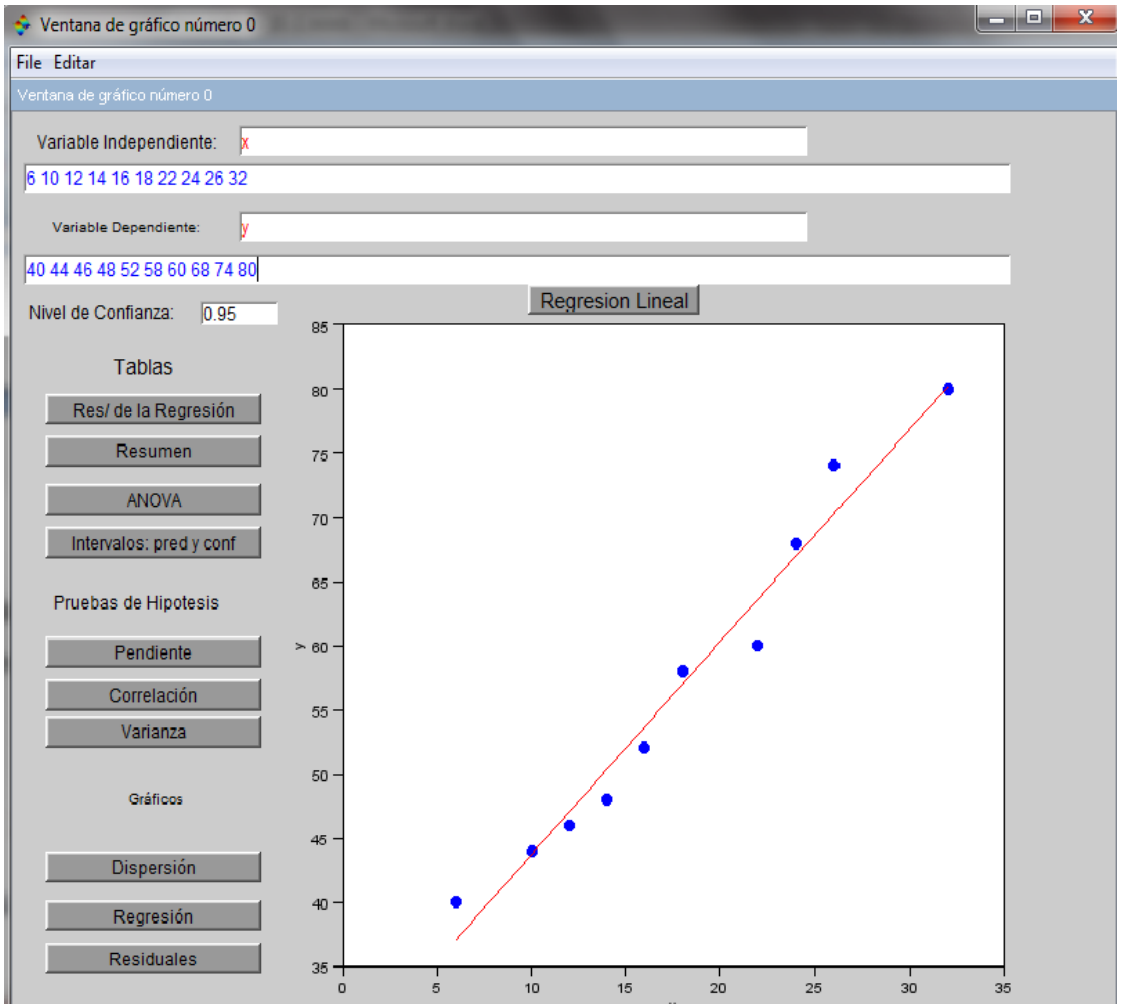
CAPÍTULO 6: ANALISIS DE RESULTADOS

1. Se elaboró una Toolbox en ambiente Scilab, llamada RegDan, para la enseñanza y aprendizaje de la regresión lineal simple y múltiple.



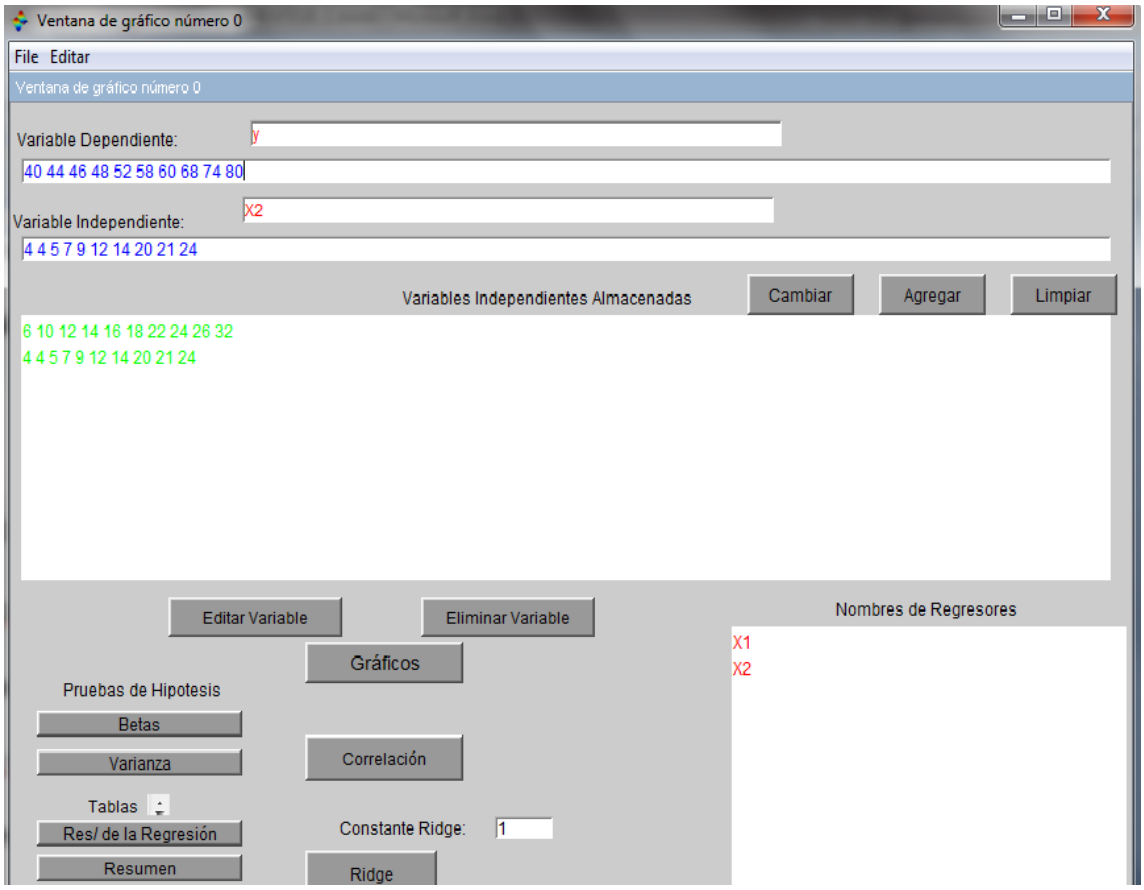
La anterior figura ilustra el menú principal de la ToolBox RegDan. Ésta posee tres opciones de trabajo. Una mediante una interfaz gráfica para regresión lineal simple, otra interfaz para regresión lineal múltiple, y un tercer modo para trabajar en la pantalla de comandos (ver anexo 1).

La interfaz gráfica para regresión simple se muestra a continuación:

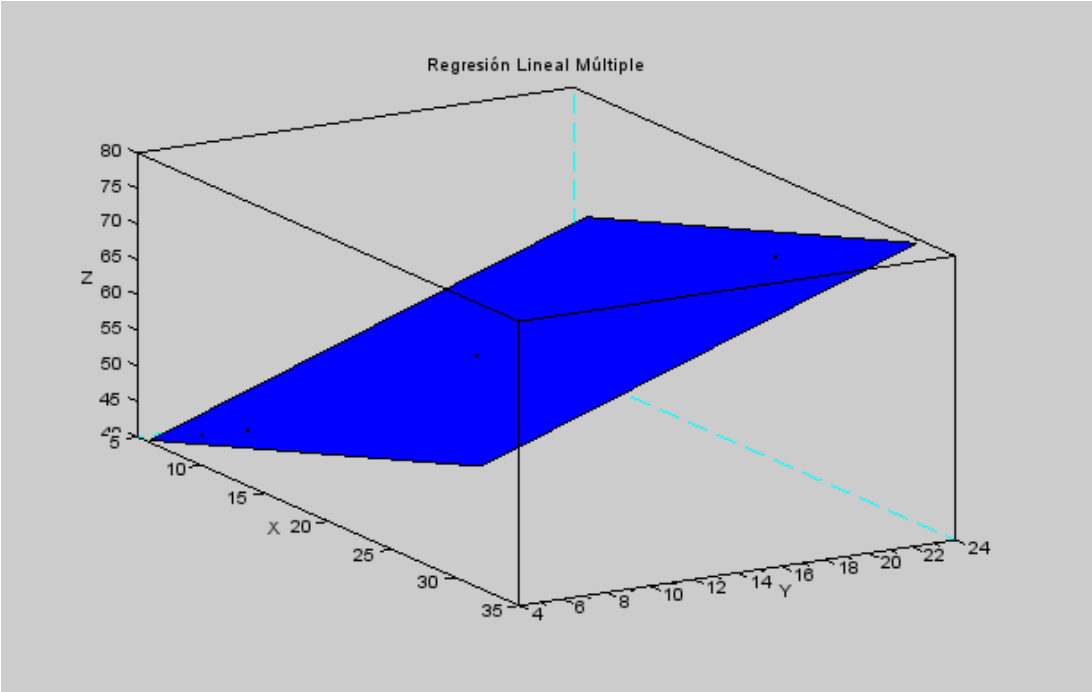


Esta interfaz permite ingresar fácilmente los datos estadísticos. Los Botones del lado izquierdo realizan automáticamente un conjunto de operaciones relacionadas con la regresión lineal simple. (Ver anexo 1)

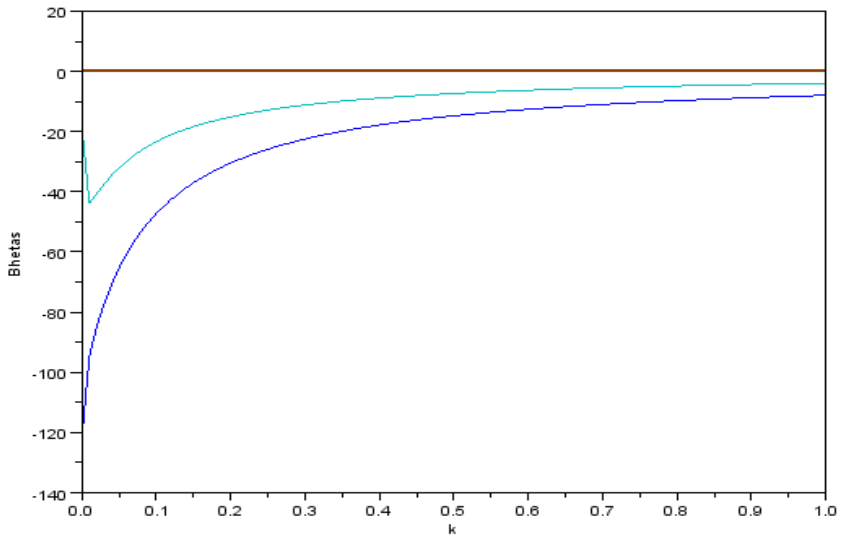
A continuación se ilustra la interfaz gráfica para la regresión múltiple:



Esta interfaz permite ingresar fácilmente los datos estadísticos. Los Botones realizan automáticamente un conjunto de operaciones relacionadas con la regresión lineal múltiple. (Ver anexo 1). Esta interfaz permite además generar gráficos en tres dimensiones: diagramas de dispersión, residuales y planos de regresión. A continuación se ilustra el plano de regresión generado por esta interfaz:



Se elaboraron rutinas basadas en el método de Ridge para la solución del problema de la multicolinealidad y para la obtención de un parámetro optimo de penalización. (ver anexo 1). A continuación se ilustra una gráfica del método de las trazas de Ridge para la obtención del k óptimo:



Esta gráfica ilustra la evolución de los parámetros de regresión a medida que la constante de Ridge aumenta de 0 a 1. La completa explicación de la gráfica y su significado se puede consultar la última parte del anexo 1.

2. Se elaboró un tutorial para el estudiante, el cual contiene la teoría de un curso de regresión lineal, con el fin de agilizar el método tradicional de enseñanza aprendizaje y poder resolver un ejercicio aplicado en el tablero de forma manual. Ver anexo 1.

3. En el aula de clase se resolvieron dos ejercicios de forma manual, uno de la recta y otro de regresión lineal, con la finalidad de recordar ciertos conceptos de la recta y para luego diferenciar estos, con los nuevos conceptos de la regresión lineal.

Luego los 28 estudiantes se llevaron a un aula de cómputo, donde con antelación ya se tenía montado el toolbox RegDan.

A cada estudiante se le hace entrega el tutorial de RegDan, y se les pide que lo desarrollen en su totalidad. Ver Anexo Nº 1.

Al final de clase a los 28 estudiantes se les pide evaluar el Toolbox mediante un formato de prueba de Likert que busca la favorabilidad o desfavorabilidad en la enseñanza de la regresión lineal por medio del tutorial. Ver Anexo Nº 2.

Los resultados de las valoraciones por parte de los estudiantes fueron las siguientes:

	VALORACIÓN					
	1	2	3	4	5	S.T
1	5	5	5	5	4	24
2	3	5	4	3	4	19
3	5	4	5	5	5	24
4	5	4	4	5	4	22
5	4	5	4	4	4	21
6	5	5	4	5	5	24
7	5	5	5	4	5	24
8	4	5	5	5	4	23

9	4	5	4	5	4	22
10	5	4	4	4	4	21
11	4	5	4	5	4	22
12	5	5	5	5	5	25
13	4	4	5	4	5	22
14	3	5	4	5	3	20
15	5	4	5	5	5	24
16	4	4	3	5	5	21
17	4	3	4	4	4	19
18	4	5	5	4	5	23
19	5	5	5	5	5	25
20	5	5	4	5	5	24
21	4	5	4	5	5	23
22	5	5	4	5	5	24
23	4	4	3	4	4	19
24	4	5	3	4	4	20
25	5	5	5	5	5	25
26	4	4	3	5	3	19
27	5	5	4	5	4	23
28	4	5	4	5	5	23

Para la valoración es indispensable tener en cuenta que el valor 1 corresponde a muy en desacuerdo, el 2 en desacuerdo, 3 indeciso, 4 de acuerdo, 5 muy de acuerdo. Ver anexo 1. El puntaje menor que se puede obtener es de 5 y el mayor de 25.

S.T es la suma total de la puntuación obtenida

4. En el departamento de física y matemáticas de la Universidad Autónoma de Manizales se le solicita a 7 docentes de diferentes áreas de matemáticas, que evaluaran el toolbox RegDan y que emitirán un concepto por medio de una prueba de Likert. Ver Anexo Nº 3.

	VALORACÓN						S.T
1	29	5	5	4	5	4	23
2	30	5	5	5	5	5	25
3	31	5	5	5	5	5	25
4	32	5	5	5	5	5	25
5	33	5	5	5	5	5	25
6	34	5	5	5	5	5	25

7	35	5	5	5	5	5	25

Los resultados obtenidos fueron excelentes, ya que el promedio nos dio 22,8 puntos con una desviación estándar de 2,026 de las 35 pruebas Likert realizadas entre alumnos y docentes de la Universidad.

5. Se elaboró un cd llamado Toolbox RegDan que contiene el Toolbox, una guía para la instalación y el tutorial y que se anexa a este trabajo.

CAPITULO 7: CONCLUSIONES

- Con los supuestos simplificadores de la regresión lineal por el método de mínimos cuadrados, se dictan las clases de regresión lineal simple y regresión lineal múltiple, pero estos supuestos en realidad se deben comprobar para lograr un adecuado uso de este método tan usado en cualquier carrera universitaria a nivel de pregrado o bien en estudios más avanzados que requieran una investigación formal y seria. Cosa que en el aula sería imposible por el factor tiempo. Por tales razones es necesario el uso de algún paquete estadístico o un programa de cómputo como el ***Scilab.***

- Debido a que los cálculos a realizar en regresión lineal son tan largos y tediosos, en este documento demostraremos como se llega por el método de mínimos cuadrados para hallar las fórmulas para calcular los parámetros $\hat{\beta}_0 \; y \; \hat{\beta}_1$, y otros estadísticos importantes, sin profundizar en muchos ya que lo que se busca es una compresión conceptual de los términos más usados en la regresión a nivel de pregrado.

- Este trabajo se puede considerar como una propuesta metodológica, ya que así, se uso por primera vez el toolbox RegDan en dos cursos de regresión lineal. También es una muestra de la forma como se empleo y se dicto el tema de regresión lineal simple en dos cursos de estadística II en la Universidad Autónoma de Manizales (Caldas, Colombia), paso a paso.

- Las ventajas e impacto que tuvo el uso de un toolbox RegDan en ambiente SCILAB, en la enseñanza aprendizaje de la regresión lineal simple y múltiple, en la práctica fueron: la reducción en el tiempo en la ejecución de las operaciones que nos permitió dedicar más tiempo a la parte conceptual, y la posibilidad que tiene el estudiante de interactuar con el toolbox para observar lo que sucede cuando se realizan cambios en los datos.

- Con el toolbox RegDan el estudiante puede observar en pantalla si una hipótesis es rechazada o aceptada, puede hallar la matriz de correlación y puede usar la regresión de Ridge cuando no se pueda usar el método de mínimos cuadrados.

CAPITULO 8: RECOMENDACIONES

- Se recomienda el uso e implementación del toolbox RegDan como una estrategia didáctica ya que se observó en los estudiantes y docentes de la Universidad una gran favorabilidad como herramienta informática que ayuda a la enseñanza aprendizaje de la regresión lineal.

- Se sugiere crear un manual que se encuentre a la mano de los usuarios cada vez que se aplique.

BIBLIOGRAFIA

[1] http://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal.

[2] http://es.wikipedia.org/wiki/Scilab.

[3] Webster, Allen L**.,** Estadistica aplicada a la empresa y a la Economia.McGraw-hill. Madrid. 1.996.

[4] Gujarati, Damodar N. Econometría. McGraw- Hill : Santafé de Bogota. 2.002

[5] Mongomery, Douglas C., Peck Elizabeth A.,Vininig, Geoffreey. Introduccion al analisis de regresión lineal. Cecsa. Mexico. 2.002

[6] Peña, Daniel. *Estadística: Modelos y Métodos.* Madrid : Alianza, 1993.

[7] Walpole, Ronald E.,Myers, Raymond H., Myers, Sharon L. Probabilidad y Estadística para ingenieros. McGraw-hill. 1.998 Mexico

[8] Salvatore, DominicK., Reagle, Derrick. Estadistica y Econometria. McGraw-hill. 2.004 España

[9] Box, George E. P., Hunter, William G., Hunter, J. Stuart. Estadistica para investigadores, introduccion al diseño de experimentos, analisis de datos y construccion de modelos. Editorial Reverté. 1989. Barcelona

[10] Cramér, Harald. Elementos de la teoria de probabilidaddes y algunas de sus alicaciones. Agilar. 1.954

[11] Freund, John E., Miller, Irwin., Miller, Marylees. Estadística Matematica con aplicaciones. Pearson. Mexico 2.000

[12] Gutierrez Pulido, Humberto., De La Vara Salazar, Roman. Analis y Diseño de Experimentos. McGraw-hill. 2.008 mexico.

[14] Hanke, John E., Reitsch, Arthur. Estadística para negocios. McGraw-hill. 1.997 Mdrid.

[15] Mendenhall, William., Scheaffer, Richard L., Wackerly, Dennis D. Estadística Matematica con Aplicaciones.

[16] http://translate.google.com.co/translate?hl=es&sl=fr&tl=es&u=http%3A%

2F%2Fwww.enpc.fr%2F

[17] http://www.google.com.co/imgres?q=grafica+de+coeficiente+de+correlacion

&hl=es&sa=X&biw=1366&bih=673&tbm=isch&prmd=imvns&tbnid=fPlNuCawHVYLpM:&imgrefurl=htp

[18] Huertas J. M., (s.f) Nuevas Tecnologías en la didáctica de la Estadística: webquest**.** http://www.uv.es/asepuma/XIV/comunica/113NUEVA.pdf

[19] Moreno Echavarría R., (2012) Propuesta didáctica para la Enseñanza de la Estadística en los modelos de Regresión Lineal bajo un enfoque constructivista. http://www.bdigital.unal.edu.co/5843/1/32561357.2012.pdf

[20] Cirilo M., Molina M., (2011) Análisis de una experiencia educativa en la modalidad B – learning. Alme (25).

http://www.clame.org.mx/documentos/alme25.pdf

[21] Molinet J., Martinez J, Casas L. (2011) Empleo de la programación en los métodos probabilísticos para la generación de números aleatorios y sus aplicaciones en la simulación. Alme (25).

http://www.clame.org.mx/documentos/alme25.pdf

[22] Contreras M., Molina E., Arteaga P. (2010) Introducción a la Estadística con R para profesores.

http://www.ugr.es/~batanero/ARTICULOS/libros/libroR.pdf

[23] Torres, G., Gibert R. (2007) La Enseñanza de la Probabilidad y la Estadística usando Statgraphics. http://www.clame.org.mx/documentos/alme20.pdf

[24] http://www.scilab.org/support/documentation/tutorials

[25] http://wiki.scilab.org/howto/Create%20a%20toolbox

[26] Ministerio de Educación Nacional. Seminario Nacional de formación de Docentes: Uso de nuevas tecnologías en el aula de matemáticas. República de Colombia 2.002

[27] Asmar Charris, Abrahan J., Topicos en teoria de matrices. Universidad Nacional de Colombia, Sede Medellín. 1.995.

ANEXOS

Anexo Nº 1

TUTORIAL EN AMBIENTE SCILAB DE REGRESIÓN LINEAL

INTRODUCCIÓN

Scilab es una aplicación matemática que dispone de un lenguaje de programación de alto nivel, para cálculo científico, interactivo de libre uso y disponible en múltiples sistemas operativos. El enfoque de Scilab es totalmente matricial, muy similar al de su homólogo comercial Matlab. Este tipo de enfoque permite lidiar fácilmente con arreglos de muchos datos, como por ejemplo los problemas estadísticos en los que se ven involucradas muchas variables y un número considerable de muestras estadísticas.

El paquete Scilab se puede descargar gratuitamente desde la página www.scilab.org. Aquí podrá encontrar además muchos tutoriales que le ayudarán a dominar totalmente el paquete.

Este tutorial está destinado a ilustrar a través de ejemplos la manera de emplear el paquete Scilab para solucionar problemas estadísticos relacionados con la regresión lineal simple y múltiple.

En este tutorial no se profundizará en los detalles de Scilab, ni en los de su lenguaje de programación. El problema se limitará a la descripción de las secuencias de código

necesarias para reproducir ciertos resultados. También se ilustrará acerca del manejo de una ToolBox (Caja de Herramientas) elaborada por el autor. Esta ToolBox hará más fácil la compresión y solución de problemas de regresión lineal.

LA LÍNEA RECTA

En el siguiente ejemplo se ilustra una distribución de puntos que yace exactamente sobre una línea recta.

Ejemplo 1

Supongamos que la estatura de 5 estudiantes de la Universidad está relacionada exactamente con la edad y sus datos están en la siguiente tabla:

Edad (años)	**Estatura (cm)**
X	**Y**
18	164
19	166
20	168
21	170
22	172

Tabla 1

Para graficar esta distribución de puntos en Scilab primero se define los vectores de datos. El vector ***x*** corresponde con la edad en años, el vector ***y*** corresponde con la estatura en centímetros. A continuación se muestra la forma de codificar estos vectores en Scilab:

```
x=[18 19 20 21 22];

y=[164 166 168 170 172];
```

Una vez definidos los vectores de datos, se procede a graficarlos. A continuación se muestran los comandos necesarios para hacerlo:

```
plot(x,y,'.');

title('Diagrama de dispersión en SCILAB');

xlabel('Edad (años), x');

ylabel('Estatura (centímetros), y');
```

La función “plot”, realiza una gráfica en dos dimensiones, para ello es necesario proporcionarle dos vectores de datos ***(x,y)*** en los cuales se encuentran las coordenadas de cada punto a graficar. La especificación ***‘.’*** (Un punto ente comillas simples), que se encuentra en la función “plot”, hace que los puntos queden aislados uno del otro. Si se omitiera esta especificación todos los puntos quedarían unidos mediante líneas rectas. Las funciones “title”, “ylabel” y “xlabel” se emplean para colocar el título de la grafica y etiquetar los ejes.

Si ejecutáramos las líneas de código que hasta ahora tenemos, se obtendría el diagrama de dispersión de la figura 1. Aquí se puede observar que todos los puntos se encuentran ubicados sobre una línea recta.

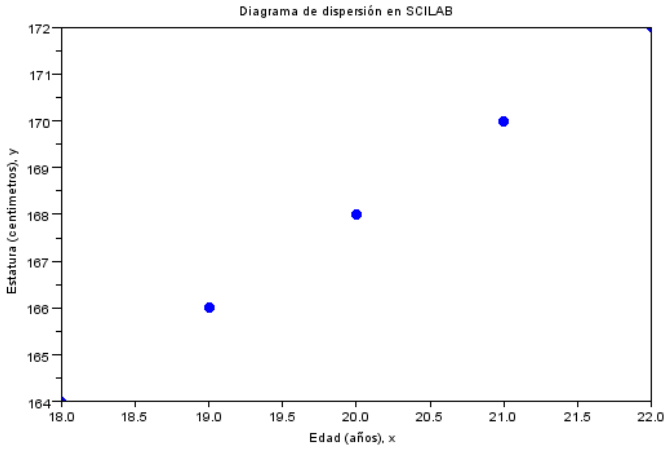


Figura 1

Con los valores de la Tabla 1 podría encontrarse analíticamente la ecuación de la recta que contiene todos los putos de esta distribución, pues se trata de una distribución de puntos idealizada, la cual coincide exactamente con una línea recta. Si quisiéramos encontrar la ecuación de la recta mediante Scilab podrimos realizar una regresión lineal de esta distribución de puntos. De esta manera sería el código:

```
[a b]=reglin(x,y);
```

La función "reglin" realiza una regresión lineal simple, para ello es necesario proporcionarle las coordenadas de los puntos ***(x,y)***. Esta función retorna dos valores: **a** y **b** los cuales corresponden a las constantes de la recta: **y = ax + b**. A continuación se muestran las líneas de código necesarias para ilustrar la recta obtenida mediante la regresión:

```
plot(x,y,'.');

plot([18 22],[a*18+b a*22+b],'r')

title('Regresión lineal en SCILAB');

xlabel('Edad (años), x');
```

```
ylabel('Estatura (centimetros), y');
```

Esta rutina superpone los puntos del diagrama de dispersión con la recta obtenida mediante la regresión. Los vales **18** y **22** corresponden a la primera y a la última edad (eje x). Los valores **a*18+b** y **a*22+b** corresponden al valor de la primera y última estatura (eje y) predichos por la ecuación de la recta **y = ax + b**. La figura 2 ilustra la ejecución de las líneas de código anteriores:

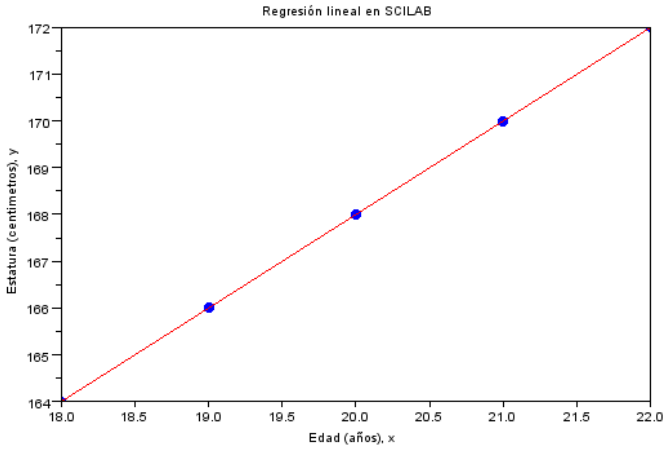


Figura 2

REGRESIÓN LINEAL SIMPLE

El siguiente ejemplo ilustra una distribución de puntos que no yace exactamente sobre una línea recta. Este ejemplo corresponde a un caso realista de muestreo estadístico en el cual se hace necesario determinar relaciones entre los datos y líneas de tendencia.

Ejemplo 2

La tabla 2, nos muestra el número de fanegas de maíz producidas por acre (Y_i), obtenidas con la utilización de diversas cantidades de fertilizante, en libras por acre (X_i), producidas en una explotación agraria en cada uno de los diez años entre 1971 y 1980.

Año	n	Y_i	X_i
1971	1	40	6
1972	2	44	10
1973	3	46	12
1974	4	48	14
1975	5	52	16
1976	6	58	18
1977	7	60	22
1978	8	68	24
1979	9	74	26
1980	10	80	32

Tabla 2

A continuación mostraremos el código fuente necesario para generar el diagrama de dispersión de este ejemplo:

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];
```

```
plot(x,y,'.');

title('Diagrama de dispersión en SCILAB');

xlabel('Fertilizante, x');

ylabel('Fanegadas de maíz, y');
```

Como en el ejemplo anterior, primero se definen los vectores de datos **x** e **y**, luego se procede a graficarlos. La ejecución de este código genera la figura 3.

A continuación se presenta la codificación necesaria para realizar la regresión lineal y graficar la recta superpuesta a los puntos:

```
figure

plot(x,y,'.');

[a b]=reglin(x,y);

plot([6 32],[a*6+b a*32+b],'r')

title('Regresión lineal en SCILAB');

xlabel('Fertilizante, x');

ylabel('Fanegadas de maíz, y');
```

El comando “figure” genera una ventana de gráficos nueva. Si no se emplea este comando todas las gráficas quedarán superpuestas en la misma ventana, a menos que no haya ventanas anteriores o ya se hayan cerrado. La figura 4 ilustra la regresión lineal.

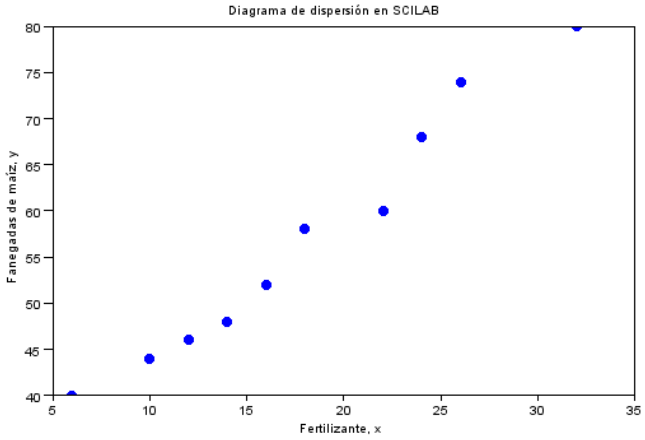


Figura 3

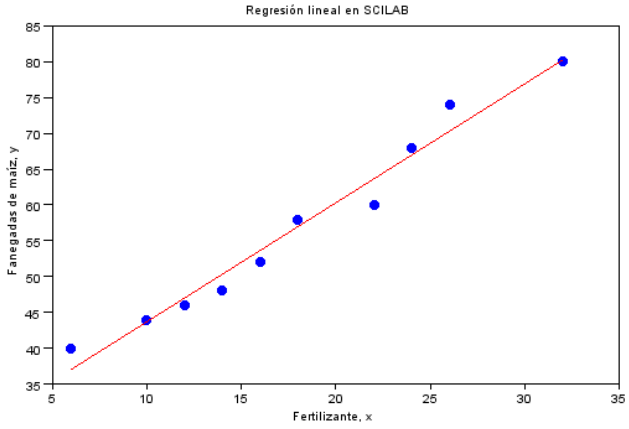


Figura 4

FORMULAS IMPORTANTES

Ahora presentaremos algunas fórmulas estadísticas relacionadas con la regresión lineal simple. Se mostrará la forma en que estas formulas podrían ser resueltas usando Scilab. Para todos los casos se usarán los datos del Ejemplo 2 (Tabla 2).

Antes de plantear cualquier fórmula, se debe definir los vectores de datos con los que se va a trabajar:

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];
```

Luego se debe determinar la cantidad de datos con la que se está trabajando. Para este caso sólo son 10, sin embargo cuando se trabaja con cantidades muy grandes, se puede utilizar la función “length” la cual retorna la cantidad de datos almacenados en un vector.

```
n=length(x);
```

Fórmula de la pendiente o coeficiente de regresión

$$b_1 = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2}$$

Scilab permite realizar operaciones con vectores de una manera muy sencilla e intuitiva. Antecediendo un punto a los operadores matemáticos (+, -,*,/) se realizan operaciones elemento a elemento entre vectores. La función “sum” realiza la suma de todos los elementos de un vector. El operador para elevar a una potencia (^), realiza la operación de la potenciación con cada uno de los elementos del vector.

Código en Scilab:

```
b1=(n*sum(x.*y)-sum(x)*sum(y))/(n*sum(x^2)-sum(x)^2)
```

Fórmula de la ordenada al origen

$$b_o = \frac{\sum y}{n} - \frac{b_1 \sum x}{n}$$

Una vez se tenga calculado el valor del parámetro `b1` se procede a calcular de manera sencilla el parámetro `b0.`

Código en Scilab:

```
b0=sum(y)/n-(b1*sum(x))/n
```

Ecuación de regresión simple

$$\hat{y} = b_0 + b_1 x$$

Ahora que se tienen calculados los valores de los dos parámetros, `b0` y `b1`, se calculan los valores estimados de *y.*

Código en Scilab:

```
ye=b0+b1*x
```

Residual

$$e = y - \hat{y}$$

Una vez calculados los valores estimados de *y* es posible calcular los residuales.

Código en Scilab:

```
e=y-ye;
```

Estimación del error estándar

$$S_{y \cdot x} = \sqrt{\frac{\sum \left(y - \hat{y}\right)^2}{n-2}}$$

Con los residuales podemos calcular la estimación del error estándar de la regresión. La función “sqrt” calcula la raíz cuadrada de un número (como en este caso) o de cada uno de los elementos de un vector.

Código en Scilab:

```
Syx=sqrt(sum(e^2)/(n-2));
```

Coeficiente de correlación de Pearson

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2}\sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

Otro valor estadístico importante es el coeficiente de correlación. Su cálculo se realiza mediante un conjunto de funciones ya explicadas anteriormente.

Código en Scilab:

```
r=(n*sum(x.*y)-sum(x)*sum(y))/(sqrt(n*sum(x^2)-
sum(x)^2)*sqrt(n*sum(y^2)-sum(y)^2));
```

Error estándar de r

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$

Una vez calculado el coeficiente de correlación, se procede a calcular de una manera muy sencilla el error estándar de éste.

Código en Scilab:

```
Sr=sqrt((1-r^2)/(n-2));
```

Ejemplo de uso matricial

En este ejemplo se ilustrará como ingresar matrices en Scilab. Para ello se resolverá un problema de regresión lineal simple desde el enfoque de solución de un problema de regresión lineal múltiple. Este enfoque requiere de matrices y operaciones entre éstas para la obtención de un vector de parámetros de regresión.

Retomando los datos de la Tabla 2, a continuación se presenta:

Año	**n**	Y_i	X_i
1971	1	40	6
1972	2	44	10
1973	3	46	12
1974	4	48	14
1975	5	52	16
1976	6	58	18
1977	7	60	22
1978	8	68	24
1979	9	74	26
1980	10	80	32

Para resolver este problema desde el enfoque de regresión lineal múltiple se emplea la siguiente fórmula matricial:

$$\beta = \left(x'x\right)^{-1} x'y$$

Donde el vector ***y***, y la matriz ***x*** se definen en la siguiente tabla:

y	***x***	
40	1	6
44	1	10
46	1	12
48	1	14
52	1	16
58	1	18
60	1	22
68	1	24
74	1	26
80	1	32

El código en Scilab para solucionar este problema es el siguiente:

```
y=[40;44;46;48;52;58;60;68;74;80];

x=[1 6;1 10;1 12;1 14;1 16;1 18;1 22;1 24;1 26;1 32];

B=inv(x'*x)*x'*y
```

Se observa que los datos del vector ***y*** se separan con punto y coma (**;**). El operador punto y coma separa una fila de otra, esto quiere decir que ***y*** es un vector columna. Hasta este momento se ha venido trabajando con vectores fila, pero para este caso que requiere matrices se hace más conveniente trabajar con vectores columna. Se observa que en la matriz ***x*** los elementos de las filas están separados por un espacio en blanco, que también podría ser una coma (**,**). Así mismo las filas están separadas una de otra por un punto y coma. La función "inv" retorna la inversa de una matriz. Si se coloca una comilla simple (**'**) a continuación de una matriz se obtendrá la transpuesta de ésta. El operador "por" (*****) realiza todo tipo de multiplicaciones, entre éstas la matricial.

TOOLBOX RegDan

RegDan es una ToolBox (Caja de herramientas) diseñada por el autor para realizar fácilmente cálculos relacionados con la regresión lineal. Esta ToolBox está dotada de un conjunto amplio de funciones que permiten calcular datos estadísticos, probar hipótesis de parámetros de regresión y realizar gráficos.

Instalación de RegDan

La caja de herramientas está contenida en una carpeta con el nombre "RegDan". Para instalarla se debe copiar esta carpeta en la carpeta "contrib" que está ubicada en el lugar donde haya quedado instalado Scilab. La ubicación de esta carpeta podría ser, por ejemplo, la siguiente: "C:\Program Files\scilab-5.3.3\contrib\RegDan"

Una vez instalada la ToolBox, ésta se podrá cargar desde Scilab. Para cargarla se debe primero abrir el programa Scilab, y luego en la pestaña "Herramientas" darle clic a la opción "RegDan". Después de esto todas las funciones de "RegDan" estarán cargadas y listas para usarse.

Funciones de la ToolBox RegDan

REGRESIÓN SIMPLE

Función

anova

Secuencia de llamado

```
f=anova(x,y)
```

Descripción: Realiza el análisis de la varianza de los datos estadísticos almacenados en los vectores ***x*** e ***y.*** Imprimen los resultados en pantalla, además retorna el valor del estadístico ***f.***

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

f=anova(x,y)
```

Función

coef_det_simple

Secuencia de llamado

```
r2=coef_det_simple(x,y)
```

Descripción: Retorna el coeficiente de determinación simple de los datos estadísticos almacenados en los vectores ***x*** e ***y.***

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];
```

```
y=[40 44 46 48 52 58 60 68 74 80];

r2=coef_det_simple(x,y)
```

Función

coeficiente_regresion

Secuencia de llamado

```
b1=coeficiente_regresion(x,y)
```

Descripción: Retorna el coeficiente de regresión de los datos estadísticos almacenados en los vectores ***x*** e ***y.***

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

b1=coeficiente_regresion(x,y)
```

Función

correlacion

Secuencia de llamado

```
r=correlacion(x,y)
```

Descripción: Retorna el coeficiente de correlación de los datos estadísticos almacenados en los vectores ***x*** e ***y.***

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

r=correlacion(x,y)
```

Función

err_est_coef_reg

Secuencia de llamado

```
Sb=err_est_coef_reg(x,y)
```

Descripción: Retorna el error estándar del coeficiente de regresión de los datos estadísticos almacenados en los vectores ***x*** e ***y.***

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

Sb=err_est_coef_reg(x,y)
```

Función

err_est_const

Secuencia de llamado

```
Sb0=err_est_const(x,y)
```

Descripción: Retorna el error estándar del la constante obtenida en la regresión. ***x*** e ***y*** son los vectores de datos estadísticos.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

Sb0=err_est_const(x,y)
```

Función

err_esta_esti

Secuencia de llamado

```
Suex=err_esta_esti(x,y,xp)
```

Descripción: Retorna el error estándar de la estimación. ***x*** e ***y*** son los vectores de datos estadísticos. ***xp*** es un valor escalar que corresponde con un valor de las datos ***x*** para el cual se desea calcular el error estándar de la estimación.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

Suex=err_esta_esti(x,y,12)
```

Función

error_estandar_pred

Secuencia de llamado

```
Syex=error_estandar_pred(x,y,xp)
```

Descripción: Retorna el error estándar de la predicción. ***x*** e ***y*** son los vectores de datos estadísticos. ***xp*** es un valor escalar que corresponde con un valor de las datos ***x*** para el cual se desea calcular el error estándar de la predicción.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

Syex=error_estandar_pred(x,y,10)
```

Función

error_estandar_r

Secuencia de llamado

```
Sr=error_estandar_r(x,y)
```

Descripción: Retorna el error estándar del coeficiente de correlación. ***x*** e ***y*** son los vectores de datos estadísticos

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

Sr=error_estandar_r(x,y)
```

Función

estimacion_error_est

Secuencia de llamado

```
Syx=estimacion_error_est(x,y)
```

Descripción: Retorna la estimación del error estándar. ***x*** e ***y*** son los vectores de datos estadísticos

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

Syx=estimacion_error_est(x,y)
```

Función

grafica_dispersion

Secuencia de llamado

```
grafica_dispersion(x,y,etiqueta_x,etiqueta_y)
```

Descripción: Realiza un diagrama de dispersión con los vectores de datos ***x*** e ***y.*** **etiqueta_x** y **etiqueta_y** son las cadenas de caracteres que corresponden a las etiquetas de los ejes ordenados.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

grafica_dispersion(x,y,'Fertilizante','Fanegadas de maíz')
```

Función

grafica_reg_lin

Secuencia de llamado

```
grafica_reg_lin(x,y,etiqueta_x,etiqueta_y)
```

Descripción: Realiza el gráfico de la recta de regresión lineal de los vectores de datos ***x*** e ***y***. **etiqueta_x** y **etiqueta_y** son las cadenas de caracteres que corresponden a las etiquetas de los ejes ordenados.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

grafica_reg_lin(x,y,'Fertilizante','Fanegadas de maíz')
```

Función

grafica_residuales

Secuencia de llamado

```
grafica_residuales(x,y,etiqueta_x,etiqueta_y)
```

Descripción: Realiza el gráfico de los residuales de los vectores de datos ***x*** e ***y***. El eje x del gráfico corresponde con la variable independiente (**x**), el eje y corresponde con los residuales. **etiqueta_x** y **etiqueta_y** son las cadenas de caracteres que corresponden a las etiquetas de los ejes ordenados.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

grafica_residuales (x,y,'Residuales','Fanegadas de maíz')
```

Función

hipotesis_correlacion

Secuencia de llamado

```
hipotesis_correlacion(x,y,confianza)
```

Descripción: Realiza la prueba de hipótesis del coeficiente de correlación. ***x*** e ***y***, son los vectores de datos estadísticos. **Confianza** es el nivel de confianza, debe ser un valor entre que esté entre 0 y 1.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

hipotesis_correlacion(x,y,0.95)
```

Función

hipotesis_pendiente

Secuencia de llamado

hipotesis_pendiente(x,y,confianza)

Descripción: Realiza la prueba de hipótesis del coeficiente de regresión. ***x*** e ***y***, son los vectores de datos estadísticos. **Confianza** es el nivel de confianza, debe ser un valor entre que esté entre 0 y 1.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];
```

hipotesis_pendiente(x,y,0.975)

Función

hipotesis_varianza

Secuencia de llamado

hipotesis_varianza(x,y,confianza)

Descripción: Realiza la prueba de hipótesis de la varianza mediante la tabla ANOVA. ***x*** e ***y***, son los vectores de datos estadísticos. **Confianza** es el nivel de confianza, debe ser un valor entre que esté entre 0 y 1.

Ejemplo

x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

hipotesis_varianza(x,y,0.99)

Función

ordenada_origen

Secuencia de llamado

b0=ordenada_origen(x,y)

Descripción: Retorna la ordenada al origen de la regresión. ***x*** e ***y*** son los vectores de datos estadísticos

Ejemplo

x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

b0=ordenada_origen(x,y)

Función

regresion_lineal

Secuencia de llamado

[b1,b0]=regresion_lineal(x,y)

Descripción: Retorna el coeficiente de regresión y la ordenada al origen. ***x*** e ***y*** son los vectores de datos estadísticos. **b1** es el coeficiente de regresión y **b2** es la ordenada al origen.

Ejemplo

x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

[b1,b0]=regresion_lineal(x,y)

Función

residual

Secuencia de llamado

```
e=residual(x,y)
```

Descripción: Retorna los residuales de la regresión de los datos ***x*** e ***y.***

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

e=residual(x,y)
```

Función

resumen_datos

Secuencia de llamado

```
resumen_datos(x,y)
```

Descripción: Presenta un cuadro resumen de los datos ***x*** e ***y***. Se muestran de los productos, cuadrados, valores estimados y residuales.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

resumen_datos(x,y)
```

Función

resumen_intervalos

Secuencia de llamado

```
resumen_intervalos(x,y,confianza)
```

Descripción: Presenta un cuadro resumen de los intervalos de confianza y predicción. ***x*** e ***y*** son los vectores de valores estadísticos. **Confianza** es el nivel de confianza, debe ser un valor entre que esté entre 0 y 1.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];
```

```
resumen_intervalos(x,y,0.95)
```

Función

resumen_resultados

Secuencia de llamado

```
resumen_resultados(x,y,E_x,E_y)
```

Descripción: Presenta un cuadro resumen de los principales resultados de la regresión. Se presenta la ecuación de la recta, un análisis de las desviaciones estándar de .los parámetros, el coeficiente de determinación y la tabla ANOVA. ***x*** e ***y*** son los vectores de valores estadísticos. **E_x** y **E_y** son las etiquetas de los vectores de datos.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

resumen_resultados(x,y,'Fertilizante','Fanegadas de maíz')
```

Función

suma_cuadrados_error

Secuencia de llamado

```
SCE=suma_cuadrados_error(x,y)
```

Descripción: Retorna la suma de cuadrados del error. ***x*** e ***y*** son los vectores de datos estadísticos.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];
```

```
y=[40 44 46 48 52 58 60 68 74 80];

SCE=suma_cuadrados_error(x,y)
```

Función

suma_cuadrados_reg

Secuencia de llamado

```
SCR=suma_cuadrados_reg(x,y)
```

Descripción: Retorna la suma de cuadrados de la regresión. ***x*** e ***y*** son los vectores de datos estadísticos.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32];

y=[40 44 46 48 52 58 60 68 74 80];

SCR=suma_cuadrados_reg(x,y)
```

Función

suma_cuadrados_total

Secuencia de llamado

```
SCT=suma_cuadrados_total(y)
```

Descripción: Retorna la suma de cuadrados de la variable dependiente. ***y*** es un vector de valores estadísticos.

Ejemplo

```
y=[40 44 46 48 52 58 60 68 74 80];
```

```
SCT=suma_cuadrados_total(y)
```

REGRESIÓN MÚLTIPLE

Función

hipotesis_varianza_m

Secuencia de llamado

```
hipotesis_varianza_m(x,y,confianza)
```

Descripción: Realiza la prueba de hipótesis de la varianza mediante la tabla ANOVA. Esta función es para regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente. **Confianza** es el nivel de confianza, debe ser un valor entre que esté entre 0 y 1.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

hipotesis_varianza_m(x,y,confianza)
```

Función

coef_det_m

Secuencia de llamado

```
r2=coef_det_m(x,y)
```

Descripción: Retorna el coeficiente de determinación para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables

independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

r2=coef_det_m(x,y)
```

Función

m_covarianza

Secuencia de llamado

```
m=m_covarianza(x,y)
```

Descripción: Retorna la matriz de covarianzas para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

m=m_covarianza(x,y)
```

Función

anova_m

Secuencia de llamado

```
f=anova_m(x,y)
```

Descripción: Realiza el análisis de varianzas para una regresión lineal múltiple y retorna el valor del estadístico **f**. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

f=anova_m(x,y)
```

Función

m_correlacion

Secuencia de llamado

```
m=m_correlacion(x,y)
```

Descripción: Retorna la matriz de correlación de los parámetros de regresión para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];
```

```
m=m_correlacion(x,y)
```

Función

cuad_medio_error

Secuencia de llamado

```
cme=cuad_medio_error(x,y)
```

Descripción: Retorna el valor del error cuadrático medio para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

cme=cuad_medio_error(x,y)
```

Función

estimacion_error_est_m

Secuencia de llamado

```
Syx=estimacion_error_est_m(x,y)
```

Descripción: Retorna el valor de la desviación estándar de la regresión para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

Syx=estimacion_error_est_m(x,y)
```

Función

suma_cuad_error_m

Secuencia de llamado

```
SCE=suma_cuad_error_m(x,y)
```

Descripción: Retorna el valor de la suma de cuadrados del error para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

SCE=suma_cuad_error_m(x,y)
```

Función

residual_m

Secuencia de llamado

```
e=residual_m(x,y)
```

Descripción: Retorna el vector de los residuales para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

e=residual_m(x,y)
```

Función

regresion_lineal_m

Secuencia de llamado

```
B=regresion_lineal_m(x,y)
```

Descripción: Retorna el vector de parámetros de regresión para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

B=regresion_lineal_m(x,y)
```

Función

m_correlacion_p

Secuencia de llamado

```
m=m_correlacion_p(x,y)
```

Descripción: Retorna la matriz de correlaciones parciales para una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

m=m_correlacion_p(x,y)
```

Función

suma_cuad_reg_m

Secuencia de llamado

```
SCR=suma_cuad_reg_m(x,y)
```

Descripción: Retorna la suma de cuadrados de una regresión lineal múltiple. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];
```

```
SCR=suma_cuad_reg_m(x,y)
```

Función

hipotesis_param_m

Secuencia de llamado

```
hipotesis_param_m(x,y,confianza)
```

Descripción: Realiza la prueba de hipótesis de los parámetros Beta. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente. ***Confianza*** es el nivel de confianza, valor entre 0 y 1.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

hipotesis_param_m(x,y,0.95)
```

Función

resumen_datos_m

Secuencia de llamado

```
resumen_datos_m(x,y)
```

Descripción: Muestra en pantalla los valores estimados de la variable dependiente y los errores. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

resumen_datos_m(x,y)
```

Función

resumen_int_m

Secuencia de llamado

```
resumen_int_m(x,y,confianza)
```

Descripción: Muestra en pantalla un resumen de los intervalos de confianza y de predicción. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente. ***Confianza*** es el nivel de confianza, valor entre 0 y 1.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

resumen_int_m(x,y,0.95)
```

Función

resumen_res_m

Secuencia de llamado

```
resumen_res_m(x,y,E_x,E_y)
```

Descripción: Muestra en pantalla los principales valores de la regresión. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente. ***E_y*** es una cadena de caracteres, es la etiqueta de la variable dependiente. ***E_x*** es un vector de cadenas de caracteres, aquí están las etiquetas de los regresores.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

resumen_res_m(x,y,['x1','x2'],'y');
```

Función

ridge

Secuencia de llamado

```
B=ridge(x,y,k)
```

Descripción: Realiza una regresión lineal múltiple por el método de Ridge, retorna los parámetros Beta. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente. ***K*** es la contante de Ridge.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];
```

```
B=ridge(x,y,0.5)
```

Función

int_conf_pred_m

Secuencia de llamado

```
int_conf_pred_m(x,y,xp,confianza)
```

Descripción: Muestra en pantalla el intervalo de confianza y de predicción para un valor del espacio de regresores ***xp***. ***x*** es una matriz donde se almacenan los valores de las variables independientes. ***y*** es un vector donde se almacenan los datos de la variable dependiente. ***Confianza*** es el nivel de confianza, valor entre 0 y 1.

Ejemplo

```
x=[6 10 12 14 16 18 22 24 26 32;4 4 5 7 9 12 14 20 21 24];

y=[40 44 46 48 52 58 60 68 74 80];

int_conf_pred_m(x,y,[6 40],0.95);
```

Ejemplos de elección de K para el método de regresión de Ridge

A continuación presentaremos un conjunto de datos estadísticos que presentan multicolinealidad. Presentaremos dos alternativas para enfrentar este problema. Los datos son los siguientes:

Y	X1	X2	X3
49.0	1300.0	7.5	0.012
50.2	1300.0	9.0	0.012
50.5	1300.0	11.0	0.0115
48.5	1300.0	13.5	0.013
47.5	1300.0	17.0	0.0135
44.5	1300.0	23.0	0.012

28.0	1200.0	5.3	0.04
31.5	1200.0	7.5	0.038
34.5	1200.0	11.0	0.032
35.0	1200.0	13.5	0.026
38.0	1200.0	17.0	0.034
38.5	1200.0	23.0	0.041
15.0	1100.0	5.3	0.084
17.0	1100.0	7.5	0.098
20.5	1100.0	11.0	0.092
29.5	1100.0	17.0	0.086

Ejemplo: Traza de Ridge.

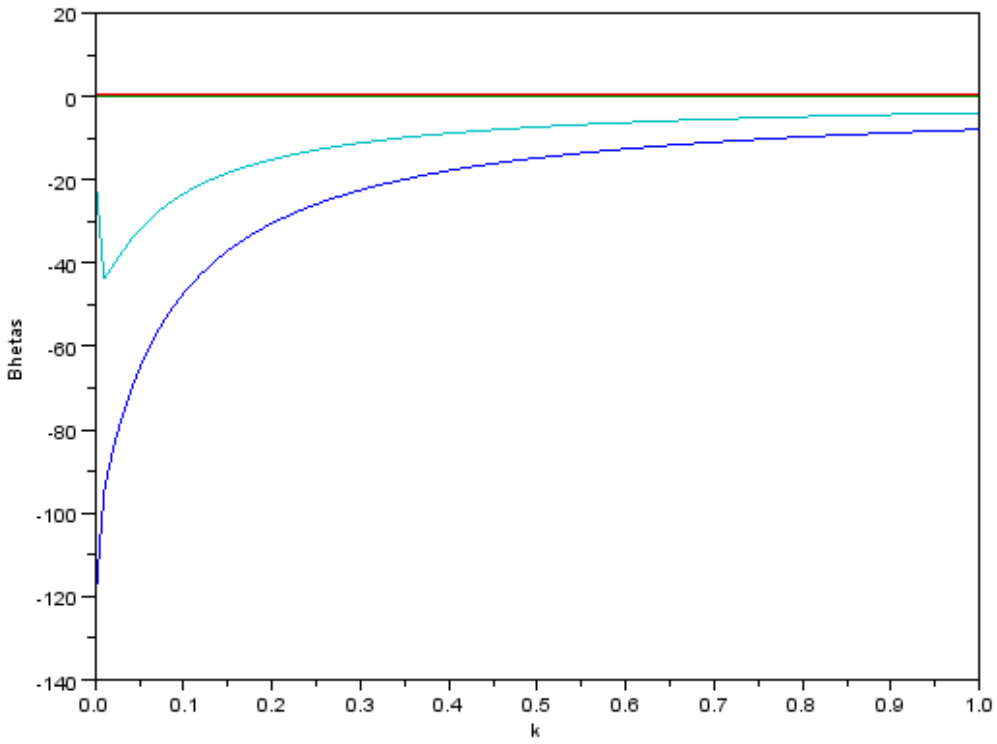
Una de las alternativas para hallar un valor adecuado de la constante de Ridge consiste en graficar la evolución de los parámetros de regresión (β) a medida que la constante de Ridge (**k**) incrementa. Generalmente la constante **k** se mueve en el intervalo [0,1]. Este tipo de análisis puede ser bastante subjetivo, pues se debe elegir un valor de **k** para el cual los parámetros β se estabilicen. Generalmente este tipo de análisis es realizado por expertos en el tema.

A continuación se muestra un trozo de código en Scilab que genera un gráfico de la traza de Ridge:

```
x=[1300 1300 1300 1300 1300 1300 1200 1200 1200 1200 1200 1200 1100 1100 1100
1100;7.5 9 11 13.5 17 23 5.3 7.5 11 13.5 17 23 5.3 7.5 11 17; 0.012 0.012 0.0115 0.013
0.0135 0.012 0.04 0.038 0.032 0.026 0.034 0.041 0.084 0.098 0.092 0.086];

y=[49 50.2 50.5 48.5 47.5 44.5 28 31.5 34.5 35 38 38.5 15 17 20.5 29.5];

    i=1;

    for k=0:0.01:1

       b=ridge(x,y,k);

       B0(i)=b(1);

       B1(i)=b(2);

       B2(i)=b(3);

       B3(i)=b(4);
```

```
    i=i+1;
end
k=0:0.01:1;
plot(k,B0,k,B1,k,B2,k,B3);
xlabel('k');
ylabel('Bhetas');
```

Al ejecutar este código se genera la siguiente gráfica:



La idea es escoger el menor **k** en donde se observe que la traza de Ridge se ha estabilizado. Un valor tentativo para **k** podría ser por ejemplo **0.4**. Sin embargo, la elección de este valor debería ser hecha por un experto en estadística y en el problema

específico que representa el modelo de regresión. Como se puede observar es muy subjetivo determinar cuando las trazas se han estabilizado.

Ejemplo: Método de Hoerl, Kennard y Baldwin

Estos tres autores y expertos en estadística sugieren que la relación que determina un valor adecuado para **k** es la siguiente:

$$k = \frac{p\hat{\delta}^2}{\hat{\beta}'\hat{\beta}}$$

Donde la varianza ($\hat{\delta}^2$) y los betas estimados ($\hat{\beta}$) son determinados por el método de los mínimos cuadrados. Basados en la metodología de superficie de respuesta en diseño experimental se determina el valor de p llamado diseño compuesto central. Este valor se usa para ajustar una superficie de respuesta de segundo orden. Se define como:

$$p = 2\gamma + C_2^{\gamma}$$

Donde γ es el número de variables regresoras, y C_2^{γ} es una combinatoria.

A continuación se muestra un trozo de código en Scilab que resuelve la anterior relación:

```
x=[1300 1300 1300 1300 1300 1300 1200 1200 1200 1200 1200 1200 1100 1100 1100
1100;7.5 9 11 13.5 17 23 5.3 7.5 11 13.5 17 23 5.3 7.5 11 17; 0.012 0.012 0.0115 0.013
0.0135 0.012 0.04 0.038 0.032 0.026 0.034 0.041 0.084 0.098 0.092 0.086];

y=[49 50.2 50.5 48.5 47.5 44.5 28 31.5 34.5 35 38 38.5 15 17 20.5 29.5];

B=regresion_lineal_m(x,y);
S=estimacion_error_est_m(x,y);

k=(9*S)/(B'*B)
```

Al ejecutar este código se obtiene un valor de **k = 0.0022500**

Anexo # 2

El objetivo de la presente encuesta es medir el grado de satisfacción del uso del programa Scilab y el toolbox RegDan utilizado en el curso de Estadística II, como recurso didáctico del **aprendizaje** de "regresión lineal".

Se presentan a continuación una serie de afirmaciones, las cuales se corresponden con cinco (5) alternativas de respuesta. Marque con una X la que considere la más adecuada.

	Muy en desacuerdo	En desacuerdo	Indeciso	De acuerdo	Muy de acuerdo
El toolbox como herramienta didáctica facilita la comprensión de la regresión lineal					
El toolbox permite ingresar fácilmente la información.					
El toolbox motiva la capacidad de análisis					
El toolbox posibilita la interacción estudiante programa					
Es factible lograr aprendizaje significativo del tema regresión lineal a través del uso del software Scilab.					

Por su valioso aporte, muchas gracias.

Anexo # 3

El objetivo de la presente encuesta es medir el grado de satisfacción del uso del programa Scilab y el toolbox RegDan utilizado en el curso de Estadística II, para la **enseñanza** de regresión lineal.

Se presentan a continuación una serie de afirmaciones, las cuales se corresponden con cinco (5) alternativas de respuesta. Marque con una X la que considere la más adecuada.

	Muy en desacuerdo	En desacuerdo	Indeciso	De acuerdo	Muy de acuerdo
El toolbox como herramienta didáctica facilita la comprensión por parte del estudiante de la regresión lineal					
El toolbox permite ingresar fácilmente la información.					
El toolbox motiva la capacidad de análisis del estudiante.					
El toolbox posibilita la interacción estudiante programa.					
Es factible lograr aprendizaje significativo del tema regresión lineal a través del uso del software Scilab.					

Por su valioso aporte, muchas gracias.